STATISTICAL METHODS FOR ENGINEERING
Probability distributions
- Continuous Random Variables -

Roberto Fontana
Department of Mathematical Sciences
Politecnico di Torino
roberto.fontana@polito.it

Data Spaces, October 2018 – January 2019

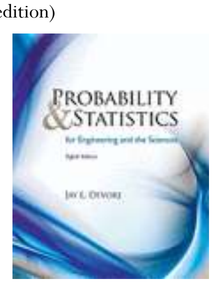
These slides are mainly based on the supporting material of Devore J., Probability and Statistics for Engineering and the Sciences (8th edition)

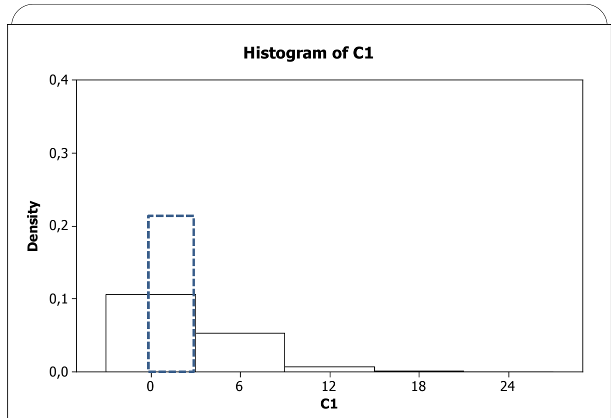
I also thank Professors Gianfranco Genta, Franco Pellerey, Suela Ruffa and Marina Santacroce for having provided me with useful material for the preparation of these notes

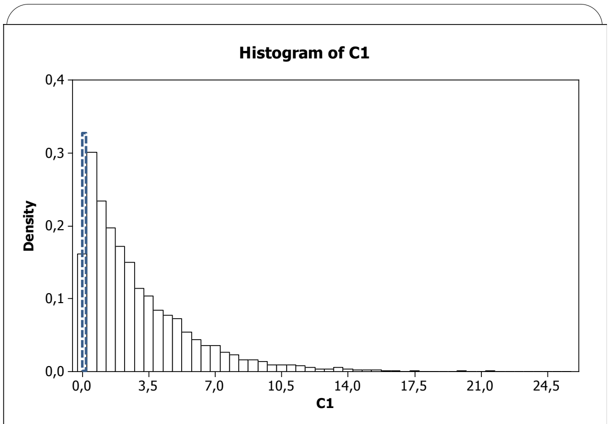
WARNING: The slides only contain brief reminders of topics; PLEASE REFER TO THE BOOK FOR STUDYING!

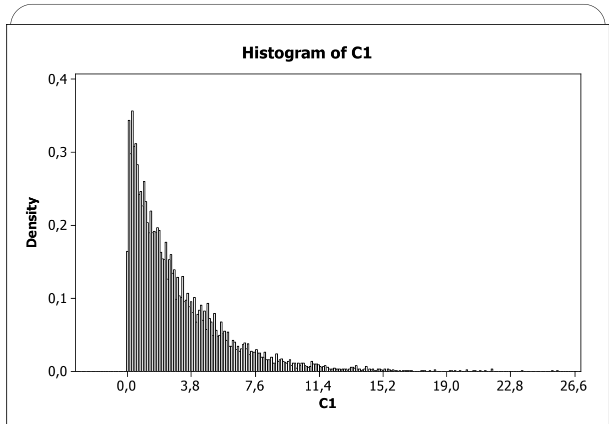
Textbook

- Devore J., Probability and Statistics for Engineering and the Sciences (8^{th} edition)









Definition and Basic Properties

Two Types of Random Variables

In Section 1.2, we distinguished between data resulting from observations on a counting variable and data obtained by observing values of a measurement variable. A slightly more formal distinction characterizes two different types of random variables.

> A **discrete** random variable is an rv whose possible values either constitute a finite set or else can be listed in an infinite sequence in which there is a first element, a second element, and so on ("countably" infinite).
>
> A random variable is **continuous** if *both* of the following apply:
>
> 1. Its set of possible values consists either of all numbers in a single interval on the number line (possibly infinite in extent, e.g., from $-\infty$ to ∞) or all numbers in a disjoint union of such intervals (e.g., $[0, 10] \cup [20, 30]$).
> 2. No possible value of the variable has positive probability, that is, $P(X = c) = 0$ for any possible value c.

→ DEV 8 p.95

Definition and Basic Properties

Recall that a random variable X is simply a function from a sample space S into the real numbers.

The random variable is discrete is the range of X is finite or countably infinite. This refers to the number of values X can take on, not the size of the values.

The random variable is (absolutely) continuous if the range of X is uncountably infinite and X has a suitable pdf (**probability density function**).

Typically an uncountably infinite range results from an X that makes a physical measurement—e.g. position, size, time, age, flow, volume, or area of something.

Definition and Basic Properties

The pdf of a continuous random variable X must satisfy three conditions.

- It is a nonnegative function (but unlike in the discrete case it may take on values exceeding 1).
- Its definite integral over the whole real line equals one. That is

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

Definition and Basic Properties

The pdf of a continuous random variable X must satisfy three conditions.

Its definite integral over a subset B of the real numbers gives the probability that X takes a value in B. That is,

$$\int_B f(x) = P(X \in B)$$

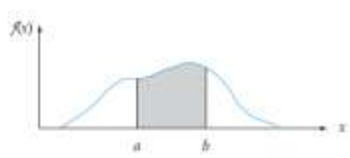
for "every" subset B of the real numbers. As a special case (the usual case) for all real numbers a and b

$$\int_a^b f(x)dx = P(a \le X \le b)$$

Put simply, the probability is simply the area under the pdf curve over the interval [a,b].

Definition and Basic Properties

If X has uncountable range and such a pdf, then X is a *continuous random variable*. In this case we often refer to f as a *continuous pdf*. Note that this means f is the pdf of a continuous random variable. It does not necessarily mean that f is a continuous function.



Definition and Basic Properties

Note that by this definition the probability of X taking on a single value a is always 0. This follows from

$$P(X = a) = P(a \le X \le a) = \int_a^a f(x)dx = 0$$

since every definite integral over a degenerate interval is 0. This is, of course, quite different from the situation for discrete random variables.

Definition and Basic Properties

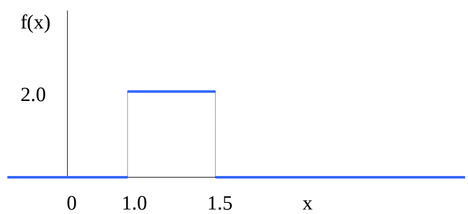
Consequently we can be sloppy about inequalities. That is

$$P(a < X < b) = P(a \le X < b) = P(a < X \le b) = P(a \le X \le b)$$

Remember that this is blatantly false for discrete random variables.

Rectangle Example: What is F(1.2)?

$F(1.2) = \Pr(X \le 1.2)$
= the area under the pdf up to where x is 1.2.

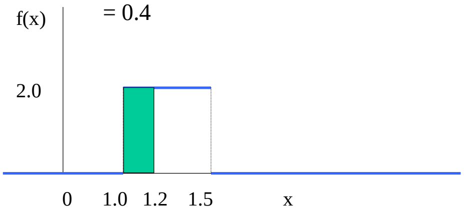


Rectangle Example: What is F(1.2)?

$F(1.2) = \Pr(X \le 1.2)$
= the area under the pdf up to where x is 1.2.

= (0.2) (2.0)

= 0.4



Example

Let X be a random variable with range [0,2] and pdf defined by f(x)=1/2 for all x between 0 and 2 and f(x)=0 for all other values of x. Note that since the integral of zero is zero we get

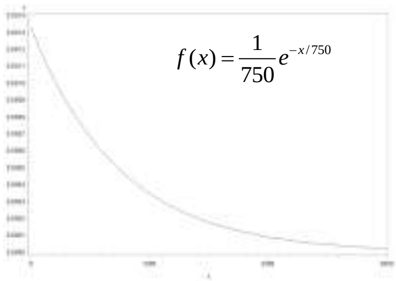
$$\int_{-\infty}^{\infty} f(x)dx = \int_0^2 1/2dx = \frac{1}{2}x\Big|_0^2 = 1 - 0 = 1$$

That is, as with all continuous pdfs, the total area under the curve is 1. We might use this random variable to model the position at which a two meters long rope breaks when put under tension, assuming "every point is equally likely". Then the probability the break occurs in the last half-meter of the rope is

$$P(3/2 \le X \le 2) = \int_{3/2}^{2} f(x)dx = \int_{3/2}^{2} 1/2dx = \frac{1}{2}x\Big|_{3/2}^{2} = 1/4$$

Example

The random variable Y might be a reasonable choice to model the lifetime in hours of a standard light bulb with average life 750 hours.



$$f(x) = \frac{1}{750} e^{-x/750}$$

Example

To find the probability a bulb lasts under 500 hours, you calculate

$$P(0 \le Y < 500) = \int_0^{500} \frac{1}{750} e^{-x/750} dx = -e^{-x/750}\Big|_0^{500} = -e^{-2/3} + 1 \approx 0.487$$

Cumulative Distribution Functions

The cdf F of a continuous random variable has the same definition as that for a discrete random variable. That is,

$$F(x) = P(X \le x)$$

In practice this means that F is essentially a particular antiderivative of the pdf since

$$F(x) = P(X \le x) = \int_{-\infty}^{x} f(t) dt$$

Thus at the points where f is continuous F'(x)=f(x).

Cumulative Distribution Functions

Knowing the cdf of a random variable greatly facilitates computation of probabilities involving that random variable since, by the Fundamental Theorem of Calculus,

$$P(a \le X \le b) = F(b) - F(a)$$

Cumulative Distribution Functions

In the second example above, F(x)=0 if x is negative and for nonnegative x we have

$$F(x)=\int_0^x \frac{1}{750}e^{-t/750}dt=-e^{-t/750}\Big|_0^x=-e^{-x/750}+1=1-e^{-x/750}$$

Thus the probability of a light bulb lasting between 500 and 1000 hours is

$$F(1000)-F(500)=(1-e^{-1000/750})-(1-e^{-500/750})=e^{-2/3}-e^{-4/3}\approx 0.250$$

Cumulative Distribution Functions

In the first example above F(x)=0 for negative x, F(x)=1 for x greater than 2 and F(x)=x/2 for x between 0 and 2 since for such x we have

$$F(x)=\int_0^x 1/2dt=\frac{1}{2}t\Big|_0^x=\frac{1}{2}x$$

Thus to find the probability the rope breaks somewhere in the first meter we calculate F(1)-F(0)=1/2-0-1/2, which is intuitively correct.

Cumulative Distribution Functions

If X is a continuous random variable, then its cdf is a continuous function. Moreover,

$$\lim_{x\to-\infty} F(x)=0$$

and

$$\lim_{x\to\infty} F(x)=1$$

Again these results are intuitive

Expectation and Variance

Definitions

The expected value of a continuous random variable X is defined by

$$E(X)=\int_{-\infty}^{\infty} xf(x)dx$$

Note the similarity to the definition for discrete random variables. Once again we often denote it by μ. As in the discrete case this integral may not converge, in which case the expectation if X is undefined.

Expectation and Variance

Definitions

As in the discrete case we define the variance by

$$Var(X) = E((X-\mu)^2)$$

Once again the standard deviation is the square root of variance. Variance and standard deviation do not exist if the expected value by which they are defined does not converge.

$$Var(X) = E(X^2) - (E(X))^2$$

$$E(X^2) = \int_{-\infty}^{\infty} x^2 f(x)dx$$

Expectation and Variance

Theorems

$$E(h(X)) = \int_{-\infty}^{\infty} h(x) f(x)dx$$

Expected value still preserves linearity. That is

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

The proof depends on the linearity of the definite integral (even an improper Riemann integral).

Expectation and Variance

Theorems

Similarly the expected value of a sum of functions of X equals the sum of the expected values of those functions by the linearity of the definite integral.

The shortcut formula for the variance holds for continuous random variables, depending only on the two preceding linearity results and a little algebra, just as in the discrete case. The formula states

$$Var(X) = E(X^2) - E(X)^2 = E(X^2) - \mu^2$$

Variance and standard deviation still act in the same way on linear functions of X. Namely

$$Var(aX+b) = a^2 Var(X)$$ and

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

Example

In the two-meter-wire problem, the expected value should be 1, intuitively. Let us calculate:

$$E(X) = \int_0^2 x\left(\frac{1}{2}\right)dx = \int_0^2 \frac{1}{4} x\,dx = \frac{1}{4}x^2\Big|_0^2 = 1-0=1$$

Example

In the same example the variance is

$$Var(X)=E(X^2)-1^2=\int_0^2 x^2\left(\frac{1}{2}\right)dx-1=\frac{1}{6}x^3\bigg|_0^2-1=\frac{1}{3}$$

and consequently

$$SD(X)=\sqrt{\frac{1}{3}}=\frac{1}{\sqrt{3}}=\frac{\sqrt{3}}{3}\approx 0.577$$

This result seems plausible.

Uniform Probability Distribution

A random variable is uniformly distributed whenever it is equally likely that a random variable could take on any value between c and d.

The uniform random variable has the following probability density function:

$$f(x)=\begin{cases}1/(b-a) & \text{for } a\le x\le b\\ 0 & \text{elsewhere}\end{cases}$$

where: a = smallest value the variable can assume

b = largest value the variable can assume

Uniform Probability Distribution

- Expected Value of X

  $E(X) = (a + b)/2$

- Variance of X

  $Var(X) = (b - a)^2/12$

Example

The service time at a restaurant is uniformly distributed between 5 and 15 minutes.

The probability density function is

$$f(x)=\begin{cases}1/10 & \text{for } 5\le x\le 15\\ 0 & \text{elsewhere}\end{cases}$$

where x is the service time for a customer

Example

What is the probability that the time that it will take to service a customer is between 12 to 15 minutes ?

Example

The area under f(x) between 12 and 15 is a rectangle.

area= 3* 0.1=0.3

Thus the probability that it takes between 12 to 15 minutes for a customer to get serviced is 0.3. Moreover, we can conclude that 30% of the customers will wait between 12 to 15 minutes for service.

Example (Expected Value and Variance)

Expected Service time

= (5 + 15)/2 = 10 minutes

Variance of Service times

$=(15-5)^2/12 = 8.33$ minutes

The Normal distribution

Density is given by

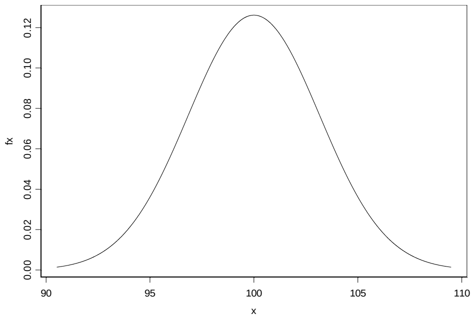
$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

μ and σ^2 are two parameters: mean and standard variance of a normal population

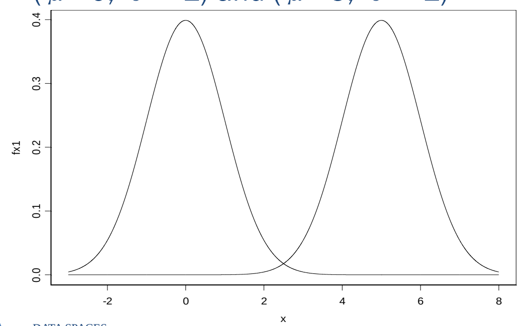
(σ is the standard deviation)

Main reason of interest: Central Limit Theorem (Lecture 5)

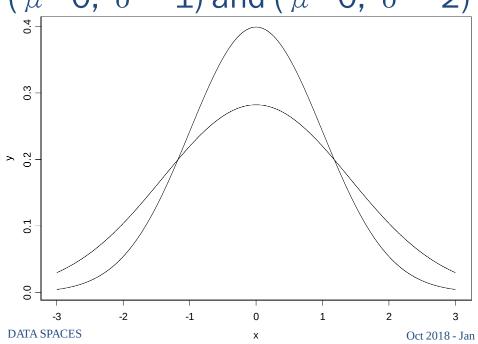
The normal — $\mu = 100$, $\sigma^2 = 10$



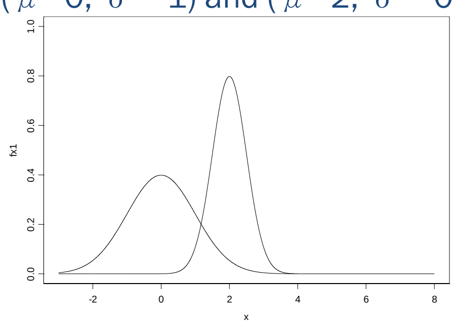
Normal curves: ($\mu = 0$, $\sigma^2 = 1$) and ($\mu = 5$, $\sigma^2 = 1$)



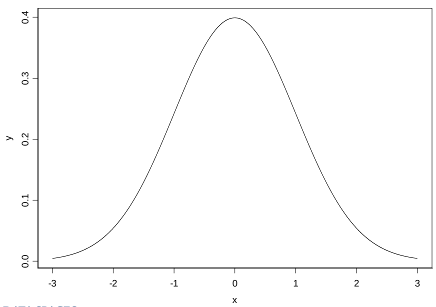
Normal curves: ($\mu = 0$, $\sigma^2 = 1$) and ($\mu = 0$, $\sigma^2 = 2$)



Normal curves: ($\mu = 0$, $\sigma^2 = 1$) and ($\mu = 2$, $\sigma^2 = 0.25$)



The standard normal curve: μ =0, and σ^2=1



How to calculate the probability of a normal random variable?

Each normal random variable, X, has a density function, say f(x) (it is a normal curve).

Probability P(a<X<b) is the area between a and b, under the normal curve f(x)

Table in the back of the book gives areas for a standard normal curve with μ=0 and σ=1.

Probabilities for any normal curve (any μ and σ) can be rewritten in terms of a standard normal curve.

Table Normal-curve Areas

Areas under standard normal curve

Areas between 0 and z (z>0)

How to get an area between a and b?

As usual, when a<b, and a, b positive

area[0,b]–area[0,a]

Probability from standard normal table

Z denotes a standard normal random variable

Standard normal curve is symmetric about the origin 0

Draw a graph

Table: P(Z<z)

z	.00	.01	.02	.03	.04	.05
0.0	.5000	.5040	.5080	.5120	.5160	.5199
0.1	.5398	.5438	.5478	.5517	.5557	.5596
0.2	.5793	.5832	.5871	.5910	.5948	.5987
0.3	.6179	.6217	.6255	.6293	.6331	.6368
0.4	.6554	.6591	.6628	.6664	.6700	.6736
0.5	.6915	.6950	.6985	.7019	.7054	.7088
0.6	.7257	.7291	.7324	.7357	.7389	.7422
0.7	.7580	.7611	.7642	.7673	.7704	.7734
0.8	.7881	.7910	.7939	.7967	.7995	.8023
0.9	.8159	.8186	.8212	.8238	.8264	.8289
1.0	.8413	.8438	.8461	.8485	.8508	.8531
1.1	.8643	.8665	.8686	.8708	.8729	.8749
1.2	.8849	.8869	.8888	.8907	.8925	.8944
1.3	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

P(Z<1,02)

Examples

- P(0<Z<1)

= P(Z<1)-P(Z<0)=0.8413-0.5=0.3413

- P(1<Z<2)

=P(Z<2)–P(Z<1)=0.9772–0.8413=0.13591

- P(Z≥1)=

=1–P(Z<1)=1–0.8413=0.1587

- P(Z ≤ -1)

=0.1587=P(Z≥1)

From non-standard normal to standard normal

X is a normal random variable with mean μ **,** and standard deviation σ

Let **Z=(X–** μ **)/** σ

Z=standard unit or z-score of X

Then **Z** has a **standard** normal distribution

$$\mu_Z = E[-\frac{\mu}{\sigma}+\frac{1}{\sigma}X] = -\frac{\mu}{\sigma}+\frac{1}{\sigma}\mu = 0 \qquad \sigma_Z^2 = Var[-\frac{\mu}{\sigma}+\frac{1}{\sigma}X] = \left(\frac{1}{\sigma}\right)^2\sigma^2 = 1$$

Example

X is a normal random variable

with μ **=120,** and σ **=15**

Find the probability P(X≤135)

Solution:

Let $Z = \frac{X-\mu}{\sigma} = \frac{X-120}{15}$

$$P(X \le 135) = P(\frac{X-120}{15} \le \frac{135-120}{15}) = P(Z \le 1) = 0.8413$$

Ex. 35 p. 162

Suppose the diameter at breast height (in.) of trees of a certain type is normally distributed with $\mu=8.8$ and $\sigma=2.8$.

a. What is the probability that the diameter of a randomly selected tree will be at least 10 in.? Will exceed 10 in.?

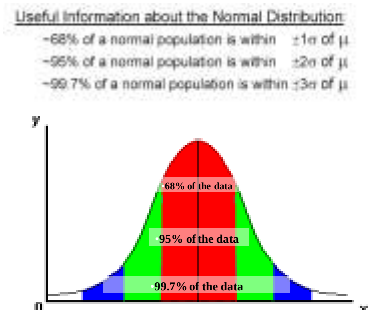
b. What is the probability that the diameter of a randomly selected tree will exceed 20 in.?

c. What is the probability that the diameter of a randomly selected tree will be between 5 and 10 in.?

d. What value c is such that the interval (8.8-c,8.8+c) includes 98% of all diameter values?

e. If four trees are independently selected, what is the probability that at least one has a diameter exceeding 10 in.?

Characteristics of Normal distribution



A useful property of the Normal Distribution 1/2

$$\text{if } X_1 \simeq \mathrm{N}\left(\mu_1, \sigma_1{}^2\right) \text{ and } X_2 \simeq \mathrm{N}\left(\mu_2, \sigma_2{}^2\right)$$

and if X_1 and X_2 are stochastically independent

then the random variable $Y = X_1 + X_2$ is such that

$$Y \simeq \mathrm{N}\left(\mu_1 + \mu_2, \sigma_1{}^2 + \sigma_2{}^2\right)$$

A useful property of the Normal Distribution 2/2

$$\text{if } X_1 \simeq \mathrm{N}\left(\mu_1, \sigma_1{}^2\right) \text{ and } X_2 \simeq \mathrm{N}\left(\mu_2, \sigma_2{}^2\right)$$

and if X_1 and X_2 are stochastically independent

then the random variable $Y = a_1 X_1 + a_2 X_2, \mathrm{a}_1, \mathrm{a}_2 \in \Re$ is such that

$$Y \simeq \mathrm{N}\left(a_1 \mu_1 + a_2 \mu_2, a_1^2 \sigma_1{}^2 + a_2^2 \sigma_2{}^2\right)$$

This property can be easily generalized to $X_1, X_1, \ldots, X_n$

→ SAS code

Exponential random variables

A nonnegative random variable X with a parameter $\lambda > 0$ obeying the following pdf and cdf is called an exponential distribution.

$$f(x)=\begin{cases}\lambda e^{-\lambda x}, \text{if } x \geq 0\\ 0, \quad \text{if } x<0\end{cases},$$

$$F(x)=P\{X \leq x\}=\int_0^x \lambda e^{-\lambda y}dy = 1-e^{-\lambda x}, x \geq 0$$

The exponential distribution is often used to describe the distribution of the amount of time until some specific event occurs.

The amount of time until an earthquake occurs; the amount of time until a telephone call you receive turns to be the wrong number

Pattern of exponential distribution

$f(x)$

$F(x)$

Expectation and variance of exponential distribution

1) $E[X]=1/\lambda$
- E[X] means the average cycle time, λ presents the occurring frequency per time interval

2) $Var(X)=1/\lambda^2$

3) The memoryless property of X
- $P\{X>s+t/X>t\}=P\{X>s\}$, if s, t≧0; $P\{X>s+t\}=P\{X>s\} \times P\{X>t\}$

$$P\{X>s+t/X>t\}=\frac{P\{X>s+t, X>t\}}{P\{X>t\}}=\frac{P\{X>s+t\}}{P\{X>t\}}$$

$$=\frac{1-F(s+t)}{1-F(t)}=\frac{1-[1-e^{-\lambda(s+t)}]}{1-[1-e^{-\lambda t}]}=e^{-\lambda s}=1-F(s)=P\{X>s\}$$

Ex. 69 p. 171

A system consists of five identical components connected in series.

As soon as one component fails, the entire system will fail.

Suppose each component has a lifetime that is exponentially distributed with $\lambda = 0.01$ and that components fail independently of one another. Define events A_i={*i*th component lasts at least *t* hours} for each i=1,…,5 so that the are independent events.

Let X the time at which the system fails—that is, the shortest (minimum) lifetime among the five components.

Ex. 69 p. 171

a. The event $\{X \geq t\}$ is equivalent to what event involving $A_1, \ldots, A_5$?

b. Using the independence of the A_i, compute $P(X \geq t)$. Then obtain $F(t)=P(X \leq t)$ and the pdf of X. What type of distribution does X have?

c. Suppose there are n components, each having exponential lifetime with parameter λ. What type of distribution does X have?

Some Exercises

Exercise 1

A certain large shipment comes with a guarantee that it contains no more than 15% defective items.

If the proportion of defective items in the shipment is greater than 15%, the shipment may be returned.

You draw a random sample of 10 items.

Let X be the number of defective items in the sample.

a. If in fact 15% of the items in the shipment are defective (so that the shipment is good, but just barely), what is $P(X \geq 7)$?
b. Based on the answer to part (a), if 15% of the items in the shipment are defective, would 7 defectives in a sample of size 10 be an unusually large number?

Exercise 1 (cont.)

c. If you found that 7 of the 10 sample items were defective, would this be convincing evidence that the shipment should be returned? Explain.
d. If in fact 15% of the items in the shipment are defective, what is $P(X \geq 2)$?
e. Based on the answer to part (d), if 15% of the items in the shipment are defective, would 2 defectives in a sample of size 10 be an unusually large number?
f. If you found that 2 of the 10 sample items were defective, would this be convincing evidence that the shipment should be returned? Explain.

Exercise 2

The number of messages received by a computer bulletin board is a Poisson random variable with a mean rate of 8 messages per hour.

a. What is the probability that 5 messages are received in a given hour?
b. What is the probability that 10 messages are received in 1.5 hours?
c. What is the probability that fewer than 3 messages are received in one-half hour?

Exercise 4

Scores on a standardized test are approximately normally distributed with a mean of 480 and a standard deviation of 90.

a. What proportion of the scores are above 700?
b. What is the 25th percentile of the scores?
c. If someone's score is 600, what percentile is she on?
d. What proportion of the scores are between 420 and 520?

Exercise 5

A cylindrical hole is drilled in a block, and a cylindrical piston is placed in the hole. The clearance is equal to one-half the difference between the diameters of the hole and the piston. The diameter of the hole is normally distributed with mean 15 cm and standard deviation 0.025 cm, and the diameter of the piston is normally distributed with mean 14.88 cm and standard deviation 0.015 cm.

a. Find the mean clearance.
b. Find the standard deviation of the clearance.
c. What is the probability that the clearance is less than 0.05 cm?
d. Find the 25th percentile of the clearance.

Exercise 5 (cont.)

e. Specifications call for the clearance to be between 0.05 and 0.09 cm. What is the probability that the clearance meets the specification?
f. It is possible to adjust the mean hole diameter. To what value should it be adjusted so as to maximize the probability that the clearance will be between 0.05 and 0.09 cm?

Supplementary material

Poisson vs. exponential

Suppose that independent events are occurring at random time points, and let N(t) denote the number of event that occurs in the time interval [0,t]. These events are said to constitute a Poisson process having rate λ, λ>0, if

- N(0)=0;
- The distribution of number of events occurring within an interval depends on the time length and not on the time point.
- $\lim_{h\to 0} P\{N(h)=1\}/h=\lambda$
- $\lim_{h\to 0} P\{N(h)\geqq 2\}/h=0$

Poisson vs. exponential (cont.)

P{N(t)=k}=P{k of the n subintervals contain exactly 1 event and the other n-k contain 0 events}

- P{exactly 1 event in a subinterval t/n}=λ(t/n)
- P{0 events in a subinterval t/n}=1-λ(t/n)

- P{N(t)=k} = $\binom{n}{k}\left(\frac{\lambda t}{n}\right)^k\left(1-\frac{\lambda t}{n}\right)^{n-k}$

binomial distribution with p=λ(t/n)

A binomial distribution approximates to a Poisson distribution with k=n(λt/n) when n is large and p is small.

$$P[N(t)=k]=e^{-\lambda t}\frac{(\lambda t)^k}{k!}, k=0,1,2,\ldots$$

Poisson vs. exponential (cont.)

Let X_1 is the time of first event. P{X_1>t}=P{N(t)=0} (0 events in the first t time length)=exp(-λt)

- F(t)=P{$X_1\leqq$t}=1- exp(-λt) with mean 1/λ
- X_1 is an exponential random variable

Let X_n is the time elapsed between (n-1)st and nth event.

P{X_n>t/X_{n-1}=s}=P{N(t)=0} =exp(-λt)

- F(t)=P{$X_n\leqq$t}=1- exp(-λt) with mean 1/λ
- X_n is also an exponential random variable

Exercise 3

A microbiologist wants to estimate the concentration of a certain type of bacterium in a wastewater sample.

She puts a 0.5 mL sample of the wastewater on a microscope slide and counts 39 bacteria.

Estimate the concentration of bacteria, per mL, in this wastewater, and find the uncertainty in the estimate.

Exercise 6

The number of traffic accidents at a certain intersection is thought to be well modeled by a Poisson process with a mean of 3 accidents per year.

a. Find the mean waiting time between accidents.
b. Find the standard deviation of the waiting times between accidents.
c. Find the probability that more than one year elapses between accidents.
d. Find the probability that less than one month elapses between accidents.
e. If no accidents have occured within the last six months, what is the probability that an accident will occur within the next year?