

Advanced Statistics

4. Parametric Families of Density Functions

Christian Aßmann

Chair of Survey Statistics and Data Analysis – Otto-Friedrich-Universität Bamberg

In this chapter we consider specific density functions of common statistical distributions, including those of binomial, Poisson, and Normal distributions.

- ▶ We usually deal with a **parametric family** of density functions. That is, the pdf f is indexed by one or more **parameters**, say θ , which allow us to vary certain characteristics of the distribution, while staying within one functional form.
- ▶ A **specific member** of a family of densities will be associated with a specific value of the parameters.
- ▶ In the following we will use the generic notation $f(x; \theta)$ to denote a family of densities for random variable X . The admissible values θ of the parameters are called the **parameter space** and will be denoted by Ω .

Each parametric family of densities has its own distinguishing characteristics that make the pdfs appropriate for specifying the probability space of some experiments and inappropriate for others.

The characteristics include

- ▶ whether the pdfs are discrete or continuous,
- ▶ whether the pdfs are restricted to positive-valued and/or integer valued random variables,
- ▶ whether the pdfs are symmetric or skewed.

In the following we will consider a list of commonly used parametric families of density functions. For each family, we will consider the major characteristics and application contexts.



Family Name: Discrete Uniform

Parameterization $N \in \Omega = \{N : N \text{ is a positive integer}\}$

Density Definition $f(x; N) = \frac{1}{N} \mathbb{I}_{\{1, 2, \dots, N\}}(x)$

Moments $\mu = (N + 1)/2, \sigma^2 = (N^2 - 1)/12, \mu_3 = 0$

MGF $M_X(t) = \sum_{j=1}^N e^{jt} / N$

Background and Application: The discrete uniform distribution assigns equal probability to each of N possible outcomes of an experiment.

Hence, the discrete uniform can be used to model the probability space of any experiment having N outcomes that are all equally likely.

EXAMPLE: Consider the experiment of rolling a die. The pdf of the number of dots facing up is

$$f(x; N = 6) = \frac{1}{6} \mathbb{I}_{\{1, 2, \dots, 6\}}(x),$$

and belongs to the family of discrete uniforms. ||

Family Name: Bernoulli

Parameterization $p \in \Omega = \{p : 0 \leq p \leq 1\}$

Density Definition $f(x; p) = p^x(1 - p)^{1-x}\mathbb{I}_{\{0,1\}}(x)$

Moments $\mu = p, \sigma^2 = p(1 - p), \mu_3 = 2p^3 - 3p^2 + p$

MGF $M_X(t) = pe^t + (1 - p)$

Background and Application: The Bernoulli distribution can be used to model experiments that have two possible outcomes often termed as *success* and *failure*, which are coded 0 (failure) and 1 (success). The event $x = 0$ has probability $f(0; p) = 1 - p$, and the event $X = 1$ has probability $f(1; p) = p$

EXAMPLE: A simple example consists of tossing a coin with a probability of a head p and $x = 1$ if head faces up. ||

The Bernoulli distribution plays an important role in **microeconometrics**, where we often consider discrete binary 0-1 decisions of consumers or households (for example the decision to buy or not to buy a certain product).

Family Name: Binomial

Parameterization $(n, p) \in \Omega = \{(n, p) : n \text{ is a positive integer, } 0 \leq p \leq 1\}$

Density Definition $f(x; n, p) = \begin{cases} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, & \text{for } x = 0, 1, 2, \dots, n \\ 0 & \text{otherwise} \end{cases}$

Moments $\mu = np, \sigma^2 = np(1-p), \mu_3 = np(1-p)(1-2p)$

MGF $M_X(t) = (1 - p + pe^t)^n$

Background and Application: The binomial density is used to model an experiment that consists of n independent repetitions of a Bernoulli-type experiment with a success probability p .

The quantity of interest x is the **total number of successes in n of such Bernoulli trials**.

The **functional form of the pdf** obtains directly from the construction of the experiment and can be derived as follows.

- ▶ Let (Z_1, \dots, Z_n) be a collection of n independent Bernoulli distributed random variables, each with $P(z_i = 1) = p$.

Then the random variable $X = \sum_{i=1}^n Z_i$ represents the number of the successful Bernoulli trials with $z_i = 1$.

- ▶ Since the Z_i s are independent, the probability of obtaining in a sequence of n trials a particular sequence of z_i outcomes with x successful Bernoulli trials and $n - x$ failures is

$$p^x (1 - p)^{n-x}.$$

- ▶ The number of different sequences of n trials that result in x successful Bernoulli trials and $n - x$ failures is

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

(That is the number of different ways of placing x outcomes with $z_i = 1$ into the n positions.)

- ▶ Since the $\binom{n}{x}$ different sequences are **mutually exclusive** and have the **same probability**, it follows that the probability for x successful trials is the sum of the probabilities for the individual sequences

$$P_X(x) = f(x; n, p) = \binom{n}{x} \cdot p^x (1 - p)^{n-x}.$$

The graph of the binomial density with $n = 5$ and $p = .3$ is given in Fig. 18.

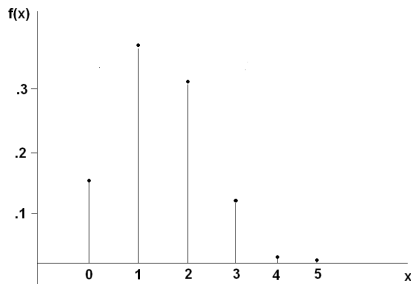


Fig. 18.

EXAMPLE: What is the probability of obtaining at least one '6' in four rolls of a fair die?

- ▶ This experiment can be modeled as a sequence of $n = 4$ Bernoulli trials with success probability $p = 1/6 = P(6 \text{ dots face up})$.
- ▶ Define the random variable $X = \text{total number of 6s in four rolls}$.
- ▶ Then $X \sim \text{binomial}(n = 4, p = 1/6)$ and

$$\begin{aligned} P(\text{at least one '6'}) &= P(X > 0) = 1 - P(X = 0) \\ &= 1 - \binom{4}{0} \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^4 = .518. \quad || \end{aligned}$$

A **generalization of the binomial distribution** to the case where there is interest in more than 2 types of outcomes for each trial is the **multinomial distribution**.

Family Name: Multinomial

Parameterization	$(n, p_1, \dots, p_m) \in \Omega = \{(n, p_1, \dots, p_m) : n \text{ is a positive integer, } 0 \leq p_i \leq 1, \forall i, \sum_{i=1}^m p_i = 1\}$
Density Definition	$f(x_1, \dots, x_m; n, p_1, \dots, p_m) = \begin{cases} \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i} & \text{for } x_i = 0, 1, 2, \dots, n \forall i, \quad \sum_{i=1}^m x_i = n \\ 0 & \text{otherwise} \end{cases}$
Moments	$\mu_i = np_i, \sigma_i^2 = np_i(1 - p_i), \mu_{3,i} = np_i(1 - p_i)(1 - 2p_i),$ $\text{Cov}(X_i, X_j) = -np_i p_j$
MGF	$M_X(t) = \left(\sum_{i=1}^m p_i e^{t_i} \right)^n$

Background and Application: The multinomial density is used to model an experiment that consists of n independent repetitions of an experiment with $m > 2$ different types of outcomes each with probability p_i .

The quantities of interest x_1, \dots, x_m are the **total numbers of each type of outcome of the experiment in n repetitions of the experiment**. That is, x_i represents the total number of outcomes of type i .

Note that the **range** of the random variable (X_1, \dots, X_n) is given by $R(X) = \{(x_1, \dots, x_n) : x_i \in \{0, 1, \dots, n\} \forall i, \sum_{i=1}^m x_i = n\}$.

The **functional form of the pdf**, given by $f(x_1, \dots, x_m; n, p_1, \dots, p_m) = \frac{n!}{\prod_{i=1}^m x_i!} \prod_{i=1}^m p_i^{x_i}$ obtains by directly extending the arguments used in the binomial case based upon recognizing

- ▶ that the probability of obtaining in a sequence of n repetitions a **particular sequence** of outcomes x_1, \dots, x_n is

$$\prod_{i=1}^m p_i^{x_i};$$

- ▶ and that the number of **different sequences** of n repetitions that result in x_i type i outcomes for $i = 1, \dots, m$ equals

$$\frac{n!}{x_1! \cdots x_m!}.$$

Family Name: Negative Binomial (Pascal)

Parameterization $(r, p) \in \Omega \{(r, p) : r \text{ is a positive integer, } 0 < p < 1\}$

Density Definition $f(x; r, p)$
$$= \begin{cases} \binom{x-1}{r-1} p^r (1-p)^{x-r} & \text{for } x = r, r+1, r+2, \dots \\ 0 & \text{otherwise} \end{cases}$$

Moments $\mu = r/p, \sigma^2 = r(1-p)/p^2, \mu_3 = r \left((1-p) + (1-p)^2 \right) / p^3$

MGF $M_X(t) = e^{rt} p^r (1 - (1-p)e^t)^{-r}$ for $t < -\ln(1-p)$

Background and Application: The negative binomial density is used to model an experiment consisting of independent Bernoulli trials with a success probability p just like the case for the binomial density.

The quantity of interest x is the **number of Bernoulli trials which are necessary to obtain r successes**.

The comparison of the binomial and the negative binomial distribution reveals that the roles of the **number of trials** and the **number of successes** are reversed w.r.t. what is a **random variable** and what is the **parameter**.

The functional form of the negative binomial pdf, given by $f(x; r, p) = \binom{x-1}{r-1} p^r (1-p)^{x-r}$, obtains from arguments similar to those used in the binomial case.

- ▶ Let (X_1, X_2, \dots) be a collection of independent Bernoulli distributed random variables, each with $P(X_i = 1) = p$.
- ▶ The probability of obtaining a particular sequence of x trials that result in r successes with the last trial being the r th success is

$$p^r (1-p)^{x-r}.$$

- ▶ The number of different sequences of x trials that result in r successes and end with the r th success is

$$\binom{x-1}{r-1} = \frac{(x-1)!}{(r-1)!(x-r)!}.$$

(That is the number of different ways of placing $r-1$ successes in the first $x-1$ positions of the sequence. Note that the last trial has to be a success.)

A special case of the negative binomial distribution is the **geometric distribution**, which obtains by setting the parameter $r = 1$. Its pdf has the form

$$f(x; p) = p(1 - p)^{x-1} \quad \text{for } x = 1, 2, \dots$$

Background and Application: The geometric density is used to model an experiment that consists of independent Bernoulli trials with a success probability p . The quantity of interest x is the **Bernoulli trial at which the first success occurs**.

The **geometric distribution** has a property known as the **memoryless property**. It means that for some positive integers i and j we obtain

$$P(x > i + j | x > i) = P(x > j).$$

That is, the probability of getting an additional j failures, having already observed i failures, is the same as the probability of observing j failures at the start of the sequence. Thus, **the geometric distribution forgets what has occurred**.

This memoryless property of the geometric distribution can be established as follows:

- First note that for any integer k

$$P(x > k) = P(\text{no success in } k \text{ trials}) = (1 - p)^k.$$

- Hence we obtain

$$\begin{aligned} P(x > i + j | x > i) &\stackrel{(\text{def.})}{=} \frac{P(x > i + j \text{ and } x > i)}{P(x > i)} = \frac{P(x > i + j)}{P(x > i)}. \\ &= \frac{(1 - p)^{i+j}}{(1 - p)^i} \\ &= (1 - p)^j = P(x > j). \end{aligned}$$

EXAMPLE: The geometric distribution is often used to model *lifetimes* of components. For example, assume a probability of $p = .001$ that a light bulb will fail on any given day.

Then the probability that the lifetime of the light bulb X will be at least 30 days is

$$P(X > 30) = \sum_{x=31}^{\infty} \underbrace{.001 \cdot (1 - .001)^{x-1}}_{\text{geometric pdf}} = .999^{30} = .970. \quad ||$$

REMARK: The memoryless property of the geometric distribution can be interpreted as a *lack-of-aging* property. It indicates that the geometric distribution is not appropriate to model lifetimes for which the failure probability is expected to increase with time. ◇

Family Name: Poisson

Parameterization $\lambda \in \Omega = \{\lambda : \lambda > 0\}$

Density Definition $f(x; \lambda) = \begin{cases} \frac{e^{-\lambda} \lambda^x}{x!}, & \text{for } x = 0, 1, 2, \dots \\ 0 & \text{otherwise} \end{cases}$

Moments $\mu = \lambda, \sigma^2 = \lambda, \mu_3 = \lambda$

MGF $M_X(t) = e^{\lambda(e^t - 1)}$

Background and Application: The Poisson density can be used to model experiments where the **quantity of interest takes values in the nonnegative integers**, for example, the number of occurrences of a certain event (such as the number of goals in a soccer game).

The shape of the Poisson density for different values of the parameter λ is illustrated in Fig. 19.

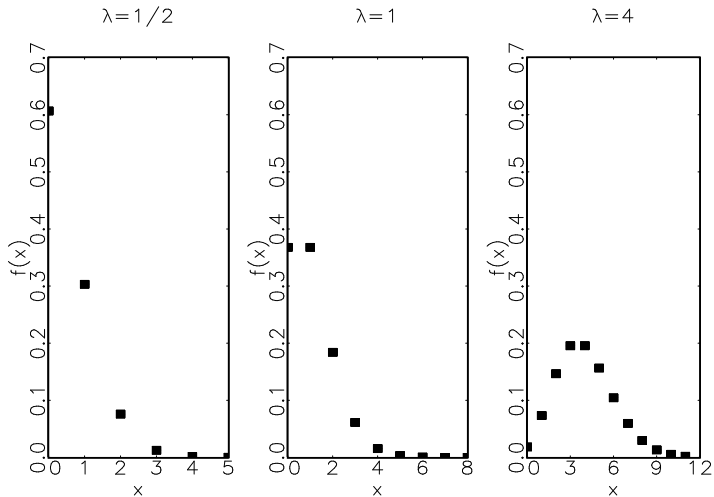


Fig. 19.

An important property of the Poisson distribution is that it provides an **approximation to the probabilities generated by the binomial distribution**. In fact, the limit of the binomial density as the number of Bernoulli trials $n \rightarrow \infty$ is the Poisson density. This can be established as follows:

The binomial density for n Bernoulli trials and success probability p is given by

$$f(x; n, p) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

- Let $np = \lambda$ for some $\lambda > 0$ such that $p = \lambda/n$. Then the binomial pdf can be rewritten as

$$\begin{aligned} f(x; n, p) &= \frac{n!}{x!(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \left[\frac{n \cdot (n-1) \cdots (n-[x-1])}{x!} \right] \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} \\ &= \left[\frac{n \cdot (n-1) \cdots (n-[x-1])}{\textcolor{blue}{n}^x} \right] \frac{\lambda^x}{x!} \left(1 - \frac{\lambda}{n}\right)^{\textcolor{blue}{n}} \left(1 - \frac{\lambda}{n}\right)^{-\textcolor{blue}{x}}. \end{aligned}$$

- Then letting $n \rightarrow \infty$, yields

$$\lim_{n \rightarrow \infty} f(x; n, p) = \lim_{n \rightarrow \infty} \left[\underbrace{\frac{n}{\textcolor{blue}{n}}}_{\rightarrow 1} \underbrace{\frac{n-1}{\textcolor{blue}{n}}}_{\rightarrow 1} \cdots \underbrace{\frac{n-x+1}{\textcolor{blue}{n}}}_{\rightarrow 1} \right] \frac{\lambda^x}{x!} \underbrace{\left(1 - \frac{\lambda}{n}\right)^n}_{\rightarrow e^{-\lambda}} \underbrace{\left(1 - \frac{\lambda}{n}\right)^{-x}}_{\rightarrow 1},$$

such that $\lim_{n \rightarrow \infty} f(x; n, p) = \frac{\lambda^x e^{-\lambda}}{x!}$, which represents the Poisson density.

- The usefulness of this result is that for a **large number of trials** n and thus for a **small success probability** $p = \lambda/n$, we can approximate the Binomial by a Poisson density, that is

$$\frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \approx \frac{(np)^x e^{-np}}{x!}.$$

Note that the Poisson density is relatively easy to evaluate, whereas, for large n , the calculation of the factorial expressions in the binomial density can be cumbersome.

Poisson density and Poisson process: The Poisson distribution models experiments whose outcomes are governed by the so-called **Poisson-process**. A Poisson process is defined as follows:

DEFINITION (POISSON PROCESS): Let an experiment consist of observing the occurrence of a certain event over a time interval $[0, t]$. The experiment is said to follow a Poisson process if:

- 1) the probability that the event occurs **once** over a small time interval Δt is approximately proportional to Δt as¹ $\gamma \cdot (\Delta t) + o(\Delta t)$, where $\gamma > 0$,
- 2) the probability that the event occurs **twice or more often** over a small time interval Δt is negligible being of order of magnitude $o(\Delta t)$,
- 3) the numbers of occurrences of the event that are observed in non-overlapping intervals are independent events.

¹ $o(\Delta t)$ stands for *of smaller order than* Δt and means that the values of $o(\Delta t)$ approach zero at a rate faster than Δt . That is $\lim_{\Delta t \rightarrow 0} \frac{o(\Delta t)}{\Delta t} = 0$.

THEOREM 4.1 *Let X be the number of times a certain event occurs in the interval $[0, t]$. If the experiment underlying X follows a Poisson process, then the pdf of X is the Poisson density.*

PROOF:

- ▶ Partition the interval $[0, t]$ into n disjoint subintervals $I_j, j = 1, \dots, n$, each of length $\Delta t = \frac{t}{n}$.
- ▶ Let $X(I_j)$ denote the number of times that the event occurs within subinterval I_j , so that $X = \sum_{i=1}^n X(I_j)$.
- ▶ Consider the event $X = k$, and note that $P(x = k) = P(A_n) + P(B_n)$, where A_n and B_n are the disjoint sets

$$\begin{aligned} A_n &= \left\{ \sum_{i=1}^n x(I_j) = k; x(I_j) \leq 1 \forall j \right\}, \\ B_n &= \left\{ \sum_{i=1}^n x(I_j) = k; x(I_j) \geq 2 \text{ for at least one } j \right\}. \end{aligned}$$

Note that $B_n \subseteq \{x(I_j) \geq 2 \text{ for at least one } j\} \subseteq \cup_{j=1}^n \{x(I_j) \geq 2\}$.

Now, Theorem 1.3 and Boole's inequality applied to $P(\cup_{j=1}^n \{x(I_j) \geq 2\})$ implies

(CONTINUES)

PROOF (CONTINUED):

$$P(B_n) \leq \underbrace{\sum_{j=1}^n P(\{x(I_j) \geq 2\})}_{\text{by property (2) of a Poisson process}} = \sum_{j=1}^n o\left(\frac{t}{n}\right) = t \left[\frac{o\left(\frac{t}{n}\right)}{\frac{t}{n}} \right].$$

It follows that $\lim_{n \rightarrow \infty} P(B_n) = 0$.

- Now consider $P(A_n)$. Define for each subinterval a **success** as observing the event once, and a **failure** otherwise. Then by property (1)

$$P(\text{success}) = \gamma \cdot \left(\frac{t}{n}\right) + o\left(\frac{t}{n}\right) \quad \text{and} \quad P(\text{failure}) = 1 - \gamma \cdot \left(\frac{t}{n}\right) - o\left(\frac{t}{n}\right)$$

- Since the occurrences of events are independent across the n subintervals by property (3), they can be interpreted as a **sequence of independent Bernoulli trials**. Hence, $P(A_n)$ obtains from a binomial distribution as

$$P(A_n) = \binom{n}{k} \left[\gamma \left(\frac{t}{n}\right) + o\left(\frac{t}{n}\right) \right]^k \left[1 - \gamma \left(\frac{t}{n}\right) - o\left(\frac{t}{n}\right) \right]^{n-k}.$$

Since $\lim_{n \rightarrow \infty} \text{Binomial}(n, p) = \text{Poisson}(\lambda = np)$, and $o(\frac{t}{n})$ disappears at a rate faster than $\frac{t}{n}$ as $n \rightarrow \infty$, we get

(CONTINUES)

PROOF (CONTINUED):

$$\lim_{n \rightarrow \infty} P(A_n) = \frac{e^{-\gamma t} (\gamma t)^k}{k!} \quad (\text{i.e. a Poisson density with } x = k),$$

where we have set $\gamma \frac{t}{n} = p$ (: Binomial parameter) such that $pn = \gamma t = \lambda$ (: Poisson parameter).

► Finally, since $P(x = k) = P(A_n) + P(B_n) \forall n$, we have

$$P(x = k) = \lim_{n \rightarrow \infty} [P(A_n) + P(B_n)] = \frac{e^{-\gamma t} (\gamma t)^k}{k!}.$$

Thus, under a Poisson process, the number of times an event occurs in an interval follows a Poisson distribution. \square

REMARK: The parameter γ is interpreted as the **mean rate of occurrence of the event per unit of time** or the **intensity of the poisson process**. This follows from the fact that for a Poisson variable $EX = \lambda = \gamma t$ such that $EX/t = \gamma$. \parallel

Family Name: Hypergeometric

Parameterization $(M, K, n) \in \Omega = \{(M, K, n) : M = 1, 2, 3, \dots; K = 0, 1, \dots, M; n = 1, 2, \dots, M\}$

Density Definition $f(x; M, K, n)$

$$= \begin{cases} \frac{\binom{K}{x} \binom{M-K}{n-x}}{\binom{M}{n}} & \text{for integer values} \\ 0 & \text{otherwise} \end{cases} \quad \max[0, n - (M - K)] \leq x \leq \min(n, K)$$

Moments $\mu = \frac{nk}{M}, \quad \sigma^2 = n \left(\frac{K}{M} \right) \left(\frac{M-K}{M} \right) \left(\frac{M-n}{M-1} \right),$

$$\mu_3 = n \left(\frac{K}{M} \right) \left(\frac{MK}{M} \right) \left(\frac{M-2K}{M} \right) \left(\frac{M-n}{M-1} \right) \left(\frac{M-2n}{M-2} \right)$$

MGF $M_X(t) = [((M-n)! (M-K)!)/M!] \times H(-n, -K, M-K-n+1, e^t),$

where $H(\cdot)$ is the hypergeometric function

$$H(\alpha, \beta, r, Z) = 1 + \frac{\alpha\beta}{r} \frac{Z}{1!} + \frac{\alpha\beta(\alpha+1)(\beta+1)}{r(r+1)} \frac{Z^2}{2!} + \dots$$

Background and Application: The hypergeometric density can be used to model experiments where there are

- 1) M objects (in an urn), of which K are of one type, say type A , and $M - K$ are of a different type, say B ;
- 2) n objects are randomly drawn from the set of the M objects *without replacement*;
- 3) the quantity of interest x is the number of type A objects in the set of n drawn objects.

Note that the **binomial** and the **hypergeometric** density **both** assign probabilities to outcomes of the type *observe x type A outcomes from a total number of n trials*.

However, under the binomial experiment, the n trials are – in contrast to the hypergeometric experiment – independent and identical and correspond to **drawing objects with replacements**.

The **functional form of the hypergeometric pdf**, given by $f(x; M, K, n) = \frac{\binom{K}{x} \binom{M-K}{n-x}}{\binom{M}{n}}$, obtains from the following arguments.

- ▶ The number of different ways of choosing the sample of size n from M objects is $\binom{M}{n}$;
- ▶ the number of different ways of choosing x type A objects is $\binom{K}{x}$;
- ▶ the number of different ways of choosing $n - x$ type B objects is $\binom{M-K}{n-x}$.

Since all possible samples having x type A objects and $n - x$ type B objects are **equally likely**, we can apply the **classical probability definition** in order to obtain the probability of drawing a sample with x type A objects

$$P(\text{sample with } x \text{ type } A \text{ objects}) = f(x; M, K, n) = \frac{\binom{K}{x} \binom{M-K}{n-x}}{\binom{M}{n}}.$$

4.2 Continuous Density Functions

Family Name: Continuous Uniform

Parameterization $(a, b) \in \Omega = \{(a, b) : -\infty < a < b < \infty\}$

Density Definition $f(x; a, b) = \frac{1}{b-a} \mathbb{I}_{[a,b]}(x)$

Moments $\mu = (a + b) / 2, \sigma^2 = (b - a)^2 / 12, \mu_3 = 0$

MGF
$$M_X(t) = \begin{cases} \frac{e^{bt} - e^{at}}{(b-a)t} & \text{for } t \neq 0 \\ 1 & \text{for } t = 0 \end{cases}$$

Background and Application: The continuous uniform density is used to model experiments having a continuous sample space with outcomes that are equally likely in the interval $[a, b]$.

EXAMPLE: A simple example consists of spinning a wheel of fortune with radius r . The point X at which the wheel stops is uniformly distributed with $a = 0$ and $b = 2\pi r$. ||

Family Name: Gamma

Parameterization $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

Density Definition $f(x; \alpha, \beta) = \frac{1}{(\beta^\alpha \Gamma(\alpha))} x^{\alpha-1} e^{-x/\beta} \mathbb{I}_{(0, \infty)}(x)$,
where $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy$ is the gamma function.

Moments $\mu = \alpha\beta, \sigma^2 = \alpha\beta^2, \mu_3 = 2\alpha\beta^3$

MGF $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \beta^{-1}$

REMARK: The gamma function has the following properties.

- ▶ $\Gamma(1) = \int_0^\infty e^{-y} dy = 1$.
- ▶ If $\alpha > 0$ is an integer, then $\Gamma(\alpha) = (\alpha - 1)!$.
- ▶ $\Gamma(1/2) = \pi^{1/2}$.
- ▶ For any real $\alpha > 0$, the gamma function satisfies the recursion $\Gamma(\alpha + 1) = \alpha\Gamma(\alpha)$. This can be verified through integration by parts:

$$\begin{aligned}
\Gamma(\alpha) &= \int_0^{\infty} \underbrace{y^{\alpha-1}}_u \underbrace{e^{-y}}_{v'} dy \\
&= \left[\underbrace{y^{\alpha-1}}_u \underbrace{(-e^{-y})}_v \right]_{y=0}^{y=\infty} - \int_0^{\infty} \underbrace{(\alpha-1)y^{\alpha-2}}_{u'} \underbrace{(-e^{-y})}_v dy \\
&= \underbrace{\lim_{y \rightarrow \infty} \frac{-y^{\alpha-1}}{e^y}}_{=0} + (\alpha-1) \underbrace{\int_0^{\infty} y^{\alpha-2} e^{-y} dy}_{\Gamma(\alpha-1)} \\
&= (\alpha-1)\Gamma(\alpha-1).
\end{aligned}$$

► The gamma function is plotted in Fig. 20. ◇

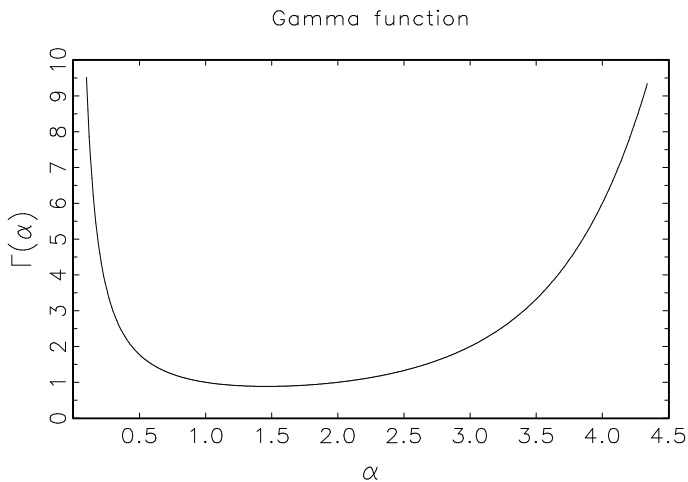


Fig. 20.

Background and Application: The gamma family of densities can be used to model experiments whose outcomes are coded as **nonnegative real numbers** and whose probabilities are to be assigned via a pdf that is **skewed to the right**.

The shape of the gamma density for different values of the parameters α and β is illustrated in Fig. 21.

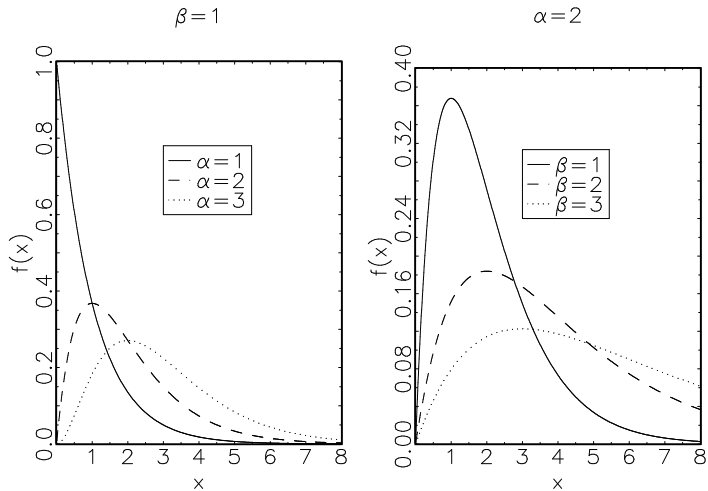


Fig. 21.

Gamma density and Poisson process: The Gamma distribution models the **waiting time (duration)** between occurrences of events under a Poisson process. This result obtains from the following arguments.

- ▶ Let Y be the **number of occurrences of an event** under a Poisson process in the interval $[0, t]$, such that $Y \sim \text{Poisson}(\lambda)$, with $\lambda = \gamma t$.
- ▶ Let X represent the **time elapsed until the first event occurs**.
- ▶ Then the probability for $y = 0$ can be written as

$$P(y = 0) = P(\text{no event in } [0, t]) = P(X > t) = \frac{e^{-\gamma t} \cdot 1}{0!}.$$

- ▶ Then the cdf for X obtains as

$$F_X(t) \stackrel{(\text{def.})}{=} P(X \leq t) = 1 - P(X > t) = 1 - e^{-\gamma t}.$$

Hence the pdf of the waiting time X can be derived as

$$\frac{\partial F_X(t)}{\partial t} = f_X(t) = \gamma e^{-\gamma t},$$

which is a gamma density with $\alpha = 1$ and $\beta = 1/\gamma$.

The gamma distribution has an **additivity property**, which is indicated in the following theorem.

THEOREM 4.2 *Let X_1, \dots, X_n be independent random variables with $X_i \sim \text{Gamma}(\alpha_i, \beta)$, $i = 1, \dots, n$. Then $Y = \sum_{i=1}^n X_i \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$.*

PROOF: Since the MGF of the X_i s are $M_{X_i}(t) = (1 - \beta t)^{-\alpha_i}$ for $t < \beta^{-1}$, and since the X_i s are independent, the MGF of Y is

$$M_Y(t) \stackrel{(\text{by indep.})}{=} \prod_{i=1}^n M_{X_i}(t) = \prod_{i=1}^n (1 - \beta t)^{-\alpha_i} = (1 - \beta t)^{-\sum_{i=1}^n \alpha_i} \quad \text{for } t < \beta^{-1}.$$

Thus, by the uniqueness theorem, $Y \sim \text{Gamma}(\sum_{i=1}^n \alpha_i, \beta)$. \square

A further property of the gamma distribution is that a **scaled gamma distributed random variable** has also a gamma distribution, as indicated in the following theorem.

THEOREM 4.3 *Let $X \sim \text{Gamma}(\alpha, \beta)$. Then, for any $c > 0$, $Y = cX \sim \text{Gamma}(\alpha, \beta c)$.*

PROOF: Since the MGF of X is $M_X(t) = (1 - \beta t)^{-\alpha}$ for $t < \beta^{-1}$, the basic properties of MGFs imply that the MGF of Y is

$$M_Y(t) = M_{cX}(t) = M_X(ct) = (1 - \beta ct)^{-\alpha} \quad \text{for } t < \beta^{-1}.$$

Thus, by the uniqueness theorem, $Y \sim \text{Gamma}(\alpha, \beta c)$. \square

An important special case of the gamma distribution, obtained by setting $\alpha = 1$ and $\beta = \theta$, is the **exponential distribution**.

Gamma Subfamily Name: Exponential

Parameterization $\theta \in \Omega = \{\theta : \theta > 0\}$

Density Definition $f(x; \theta) = \frac{1}{\theta} e^{-x/\theta} \mathbb{I}_{(0, \infty)}(x)$

Moments $\mu = \theta, \sigma^2 = \theta^2, \mu_3 = 2\theta^3$

MGF $M_X(t) = (1 - \theta t)^{-1}$ for $t < \theta^{-1}$

Background and Application The exponential density can be used to model experiments whose outcomes are coded as **nonnegative real numbers** and whose probabilities are to be assigned via a pdf that is **monotonically decreasing in x** .

A specific application of the exponential distribution is the modeling of the **time that passes until a Poisson process produces the first success** (recall the previous discussion of the relationship between the gamma density and the Poisson process).

The exponential distribution has the **memoryless property**, such that it is an appealing candidate to model **operating lives until failure** of certain objects.

THEOREM 4.4 If $X \sim \text{Exponential}(\theta)$, then $P(x > s + t | x > s) = P(x > t) \forall (t, s) > 0$.

PROOF:

$$\begin{aligned} P(x > s + t | x > s) &= \frac{P(x > s + t)}{P(x > s)} = \frac{\int_{s+t}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx}{\int_s^{\infty} \frac{1}{\theta} e^{-x/\theta} dx} \\ &= \frac{e^{-(s+t)/\theta}}{e^{-s/\theta}} = e^{-t/\theta} = P(x > t). \quad \square \end{aligned}$$

REMARK: The memoryless property implies that **given** that an object has already functioned for s units of time without failing, the probability that it will function for at least an **additional** t units of time, that is $P(x > s + t | x > s)$, is the same as the unconditional probability that it would function for at least t units of time, that is $P(x > t)$.

This indicates that the exponential distribution is not appropriate to model lifetimes for which the failure probability is expected to increase with time. \diamond

A further important special case of the gamma distribution, obtained by setting $\alpha = \nu/2$ and $\beta = 2$, is the **chi-square distribution**.

Gamma Subfamily Name: Chi-Square

Parameterization $\nu \in \Omega = \{\nu : \nu \text{ is a positive integer}\}$
 ν is called the **degrees of freedom**

Density Definition $f(x; \nu) = \frac{1}{2^{\nu/2} \Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} \mathbb{I}_{(0, \infty)}(x)$

Moments $\mu = \nu, \sigma^2 = 2\nu, \mu_3 = 8\nu$

MGF $M_X(t) = (1 - 2t)^{-\nu/2}$ for $t < \frac{1}{2}$

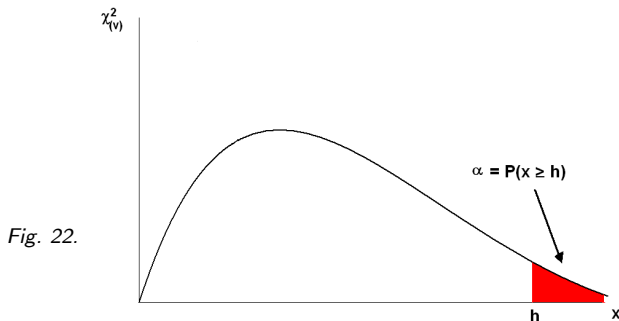
Background and Application The chi-square distribution plays an important role in **statistical inference**, especially when sampling from a normal distribution.

In particular, (as we will show later) the **sum of the squares of ν independent standard normal random variables** has a $\chi^2_{(\nu)}$ -distribution.

Furthermore, the *critical values* of many statistical tests are obtained as a quantile of a $\chi^2_{(v)}$ -distribution, that is, a value h for which

$$P(x \geq h) = \int_h^{\infty} \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2} dx = \alpha$$

(see Fig. 22).



Family Name: Beta

Parameterization $(\alpha, \beta) \in \Omega = \{(\alpha, \beta) : \alpha > 0, \beta > 0\}$

Density Definition $f(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1} \mathbb{I}_{(0,1)}(x),$

where $B(\alpha, \beta) = \int_0^1 x^{\alpha-1} (1-x)^{\beta-1} dx$ is the beta function².

Moments $\mu = \alpha/(\alpha + \beta), \quad \sigma^2 = \alpha\beta / [(\alpha + \beta + 1)(\alpha + \beta)^2],$
 $\mu_3 = 2(\beta - \alpha)(\alpha\beta) / [(\alpha + \beta + 2)(\alpha + \beta + 1)(\alpha + \beta)^3]$

MGF $M_X(t) = \sum_{r=1}^{\infty} (B(r + \alpha, \beta) / B(\alpha, \beta)) (t^r / r!)$

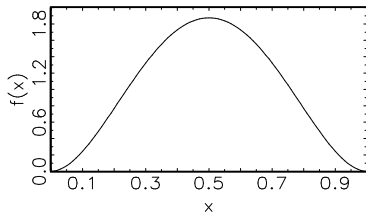
Background and Application The beta density can be used to model experiments whose outcomes are coded as [real numbers on the interval \[0, 1\]](#). It has obvious applications in modeling random variables representing [proportions](#).

²Some useful properties of the beta function include the fact that $B(\alpha, \beta) = B(\beta, \alpha)$ and $B(\alpha, \beta) = \Gamma(\alpha)\Gamma(\beta)/\Gamma(\alpha + \beta)$, so that the beta function can be evaluated in terms of the gamma function.

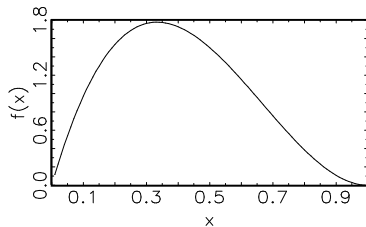
The beta density can assume a large variety of shapes depending on the parameters α and β (see also Fig. 23):

Parameter values	Shape of the beta density
$\alpha < \beta$	skewed to the right, $\mu_3 > 0$
$\alpha > \beta$	skewed to the left, $\mu_3 < 0$
$\alpha = \beta$	symmetric about the mean μ
$\alpha > 1$ and $\beta > 1$	inverted U-shaped with $\lim_{x \rightarrow 1} f(x) = 0$ and $\lim_{x \rightarrow 0} f(x) = 0$
$\alpha < 1$	$\lim_{x \rightarrow 0} f(x) = \infty$
$\beta < 1$	$\lim_{x \rightarrow 1} f(x) = \infty$
$\alpha < 1$ and $\beta < 1$	U-shaped with $\lim_{x \rightarrow 1} f(x) = \infty$ and $\lim_{x \rightarrow 0} f(x) = \infty$
$\alpha = 1$ and $\beta = 1$	uniform on $(0, 1)$

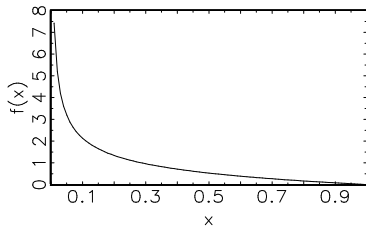
$$\alpha=3, \beta=3$$



$$\alpha=2, \beta=3$$



$$\alpha=1/2, \beta=2$$



$$\alpha=1/2, \beta=1/2$$

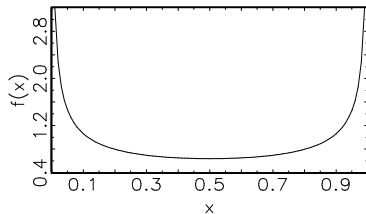


Fig. 23.

The **normal (Gaussian) family** of distributions is the most extensively used distribution in statistics and econometrics. There are three main reasons for this.

- 1) The normal distribution is very **tractable analytically**.
- 2) The normal density has a **bell shape**, whose symmetry makes it an appealing candidate to model the probability space of many experiments.
- 3) There is the **Central Limit Theorem** (which we will discuss in Chapter 5), which indicates that under mild conditions, the normal distribution can be used to approximate a large variety of distributions in large samples.

Family Name: Univariate Normal

Parameterization $(\mu, \sigma) \in \Omega = \{(\mu, \sigma) : \mu \in (-\infty, \infty), \sigma > 0\}$

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

Moments $EX = \mu, \quad \text{var}(X) = \sigma^2, \quad \mu_3 = 0$

MGF $M_X(t) = \exp \left\{ \mu t + \frac{1}{2} \sigma^2 t^2 \right\}$

Background and Application The normal family of densities is indexed by the two parameters μ and σ which correspond to the **mean** and the **standard deviation**, respectively.

In order to denote a normally distributed random variable with mean μ and variance σ^2 , we will use the usual notation $X \sim N(\mu, \sigma^2)$.

A normal distribution with $\mu = 0$ and $\sigma^2 = 1$ is called **standard normal distribution**, and is abbreviated by $N(0, 1)$.

The **functional form of the MGF** of the normal distribution obtains as follows:



$$\begin{aligned}
 M_X(t) &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} dx && \text{(by def.)} \\
 &= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{t(x-\mu) - \frac{1}{2\sigma^2}(x-\mu)^2} dx && \text{(expansion by } e^{t\mu - t\mu} \text{.)} \\
 &= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}[(x-\mu)^2 - 2\sigma^2 t(x-\mu) + \sigma^4 t^2 - \sigma^4 t^2]} dx \\
 &&& \text{(expansion by } e^{\frac{\sigma^4 t^2}{2\sigma^2} - \frac{\sigma^4 t^2}{2\sigma^2}} \text{)} \\
 &= e^{t\mu} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}[(x-\mu-\sigma^2 t)^2 - \underbrace{\sigma^4 t^2}_{\text{constant}}]} dx \\
 &= e^{t\mu + \frac{1}{2}\sigma^2 t^2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}[(x - [\mu + \sigma^2 t])^2]} dx}_{\text{pdf of a } N(\mu + \sigma^2 t, \sigma^2) \text{ which integrates to 1}} \\
 &= e^{t\mu + \frac{1}{2}\sigma^2 t^2}.
 \end{aligned}$$

The normal density is symmetric about its mean μ , has its maximum at $x = \mu$ and inflection points (where the curve changes from concave to convex) at $x = \mu \pm \sigma$ (see Fig. 24).

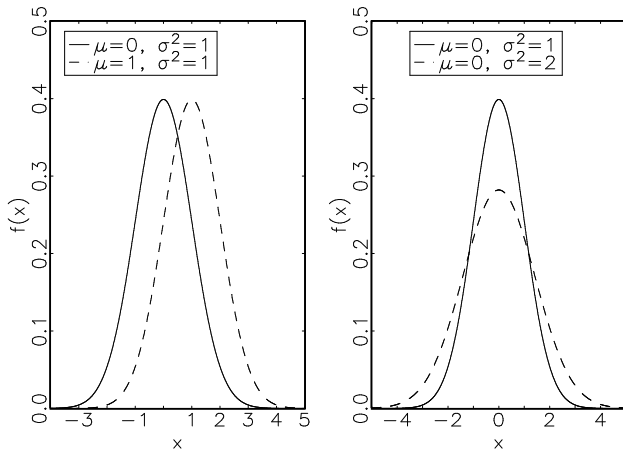


Fig. 24.

A useful property of normally distributed random variables is that they can easily be transformed into a variable having a standard normal distribution.

THEOREM 4.5 If $X \sim N(\mu, \sigma^2)$, then $Z = (X - \mu)/\sigma \sim N(0, 1)$.

PROOF: Since $-\frac{\mu}{\sigma}$ is a constant, the MGF of $Z = -\frac{\mu}{\sigma} + \frac{1}{\sigma}X$ is



$$\begin{aligned}
 M_Z(t) &= e^{-\frac{\mu}{\sigma}t} \cdot \underbrace{M_X\left(\frac{1}{\sigma}t\right)}_{\text{MGF of a } N(\mu, \sigma^2) \text{ at } t^* = t/\sigma} = e^{-\frac{\mu}{\sigma}t} \cdot e^{\mu(\frac{1}{\sigma}t) + \frac{1}{2}\sigma^2(\frac{1}{\sigma}t)^2} \\
 &= e^{\frac{1}{2}t^2},
 \end{aligned}$$

where $e^{\frac{1}{2}t^2}$ represents the MGF of a normal density with $\mu = 0$ and $\sigma^2 = 1$. \square

REMARK : Theorem 4.5 implies that the probability of an event A , $P_X(A)$, for the random variable $X \sim N(\mu, \sigma^2)$ is equal to the probability $P_Z(B)$ of the equivalent event $B = \{z : z = (x - \mu)/\sigma, x \in A\}$ for a standard normal random variable $Z \sim N(0, 1)$.

Hence, the standard normal distribution is sufficient to assign probabilities to **all** events involving Gaussian random variables. \diamond

EXAMPLE: Let $X \sim N(17, \frac{1}{4})$. The probability of the event $x \in [16, 18]$ can be computed as

$$\begin{aligned} P(16 \leq x \leq 18) &= P\left(\frac{16 - 17}{(1/2)} \leq \frac{x - 17}{(1/2)} \leq \frac{18 - 17}{(1/2)}\right) = P(-2 \leq z \leq 2) \\ &= \Phi(2) - \Phi(-2) = 0.9544, \end{aligned}$$

where $\Phi(\cdot)$ denotes the cdf of a standard normal distribution. \parallel

Normal and chi-square distribution: There is relationship between standard normal random variables and the χ^2 distribution which is subject of the following two theorems:

THEOREM 4.6 If $X \sim N(0, 1)$, then $Y = X^2 \sim \chi^2_{(1)}$.

PROOF: The MGF of Y is defined as

$$\begin{aligned}
 M_Y(t) &= E_y e^{Yt} = E_x e^{X^2 t} = \int_{-\infty}^{\infty} e^{x^2 t} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2} dx \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2(1-2t)} dx \\
 &= (1-2t)^{-1/2} \underbrace{\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}(1-2t)^{-1/2}} e^{-\frac{1}{2}\left\{\frac{x}{(1-2t)^{-1/2}}\right\}^2} dx}_{\text{pdf of a } N(0, [1-2t]^{-1}) \text{ which integrates to } 1} \\
 &= (1-2t)^{-1/2},
 \end{aligned}$$

where $(1-2t)^{-1/2}$ represents the MGF of a $\chi^2_{(1)}$ density. \square

THEOREM 4.7 *Let (X_1, \dots, X_n) independent $N(0, 1)$ -distributed random variables. Then $Y = \sum_{i=1}^n X_i^2 \sim \chi_{(n)}^2$.*

PROOF: The X_i^2 's are $\chi_{(1)}^2$ distributed by Theorem 4.6, and are independent by Theorem 2.9. Thus the MGF of Y obtains as

$$M_Y(t) \stackrel{(\text{by indep.})}{=} \prod_{i=1}^n M_{X_i^2}(t) = \prod_{i=1}^n \underbrace{(1 - 2t)^{-1/2}}_{\text{MGF of a } \chi_{(1)}^2} = (1 - 2t)^{-n/2},$$

where $(1 - 2t)^{-n/2}$ represents the MGF of a $\chi_{(n)}^2$ density. \square

The univariate normal distribution discussed so far has a straightforward multivariate generalization.



Family Name: Multivariate Normal

Parameterization $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n)'$ and $\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \cdots & \sigma_{1n} \\ \vdots & \ddots & \vdots \\ \sigma_{n1} & \cdots & \sigma_n^2 \end{pmatrix}$

$$(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \in \Omega = \{(\boldsymbol{\mu}, \boldsymbol{\Sigma}) : \boldsymbol{\mu} \in \mathbb{R}^n,$$

$\boldsymbol{\Sigma}$ is a $(n \times n)$ p.d. matrix}

Density Definition $f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$

Moments $E\mathbf{X} = \boldsymbol{\mu}, \quad \text{Cov}(\mathbf{X}) = \boldsymbol{\Sigma}, \quad \underset{(n \times 1)}{\boldsymbol{\mu}_3} = [\mathbf{0}]$

MGF $M_{\mathbf{X}}(\mathbf{t}) = \exp\{\boldsymbol{\mu}'\mathbf{t} + (1/2)\mathbf{t}'\boldsymbol{\Sigma}\mathbf{t}\}, \text{ where } \mathbf{t} = (t_1, \dots, t_n)'$.

Background and Application The n -variate normal family of distribution is indexed by $n + n(n + 1)/2$ parameters: In the mean vector (μ) n parameters and in the covariance matrix (Σ) $n + (n^2 - n)/2$ parameters.

In order to illustrate graphically some of the characteristics of a multivariate Gaussian density, we consider the bivariate case with $n = 2$.

- ▶ The multivariate Gaussian density is bell-shaped and has its maximum at $\mathbf{x} = (x_1, x_2) = \mu = (\mu_1, \mu_2)$ (see Fig. 25).
- ▶ The iso-density contours, given by the set of points $(x_1, x_2) \in \{(x_1, x_2) : f(\mathbf{x}; \mu, \Sigma) = c\}$, have the form of an ellipse. Its origin is given by μ and its direction depends on Σ (see Fig. 26).

$$\mu_x=0, \sigma_x^2=2, \mu_y=0, \sigma_y^2=1, \rho=0.5$$

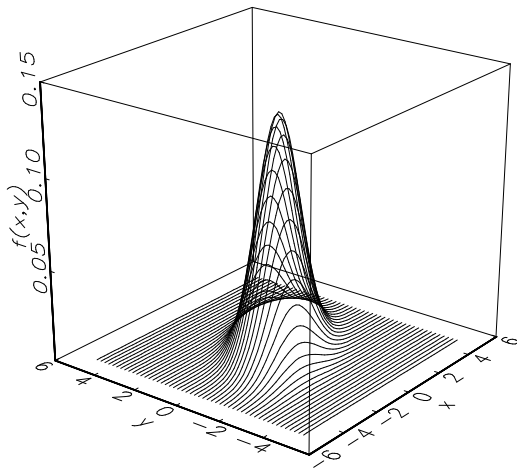


Fig. 25.

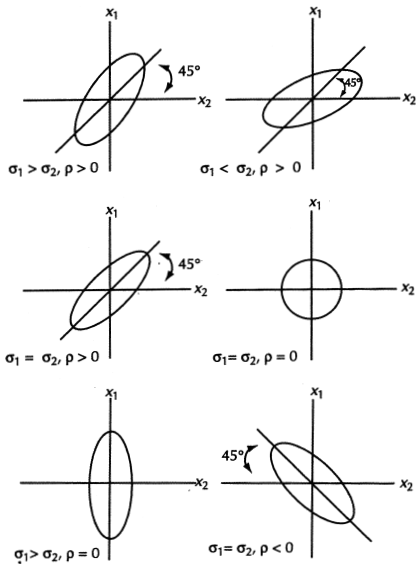



Fig. 26 (Source: Mittelhammer 1996, Fig. 4-15).

Properties of Multivariate Normal Distributions A useful property is that **linear combinations** of a vector of multivariate normally distributed random variables are also normally distributed as stated in the following theorem. 

THEOREM 4.8 *Let \mathbf{X} be an n -dimensional $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed random variable. Let \mathbf{A} be any $(k \times n)$ matrix of constants with $\text{rk}(\mathbf{A}) = k$, and let \mathbf{b} be any $(k \times 1)$ vector of constants. Then the $(k \times 1)$ random vector $\mathbf{Y} = \mathbf{AX} + \mathbf{b}$ is $N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ distributed.*

PROOF: The MGF of \mathbf{Y} is defined as

$$M_{\mathbf{Y}}(\mathbf{t}) = \mathbb{E}_{\mathbf{Y}} e^{\mathbf{t}'\mathbf{Y}} = \mathbb{E}_{\mathbf{X}} e^{\mathbf{t}'(\mathbf{AX} + \mathbf{b})} = e^{\mathbf{t}'\mathbf{b}} \cdot \mathbb{E}_{\mathbf{X}} e^{\mathbf{t}'\mathbf{A}\mathbf{X}}.$$

Defining $\mathbf{t}'\mathbf{A} = \mathbf{t}'_*$ with $\mathbf{A}'\mathbf{t} = \mathbf{t}_*$ allows us to write

$$\begin{aligned} M_{\mathbf{Y}}(\mathbf{t}) &= e^{\mathbf{t}'\mathbf{b}} \cdot \mathbb{E}_{\mathbf{X}} e^{\mathbf{t}'_*\mathbf{X}} = e^{\mathbf{t}'\mathbf{b}} \cdot M_{\mathbf{X}}(\mathbf{t}_*) = e^{\mathbf{t}'\mathbf{b}} \cdot e^{\boldsymbol{\mu}'\mathbf{t}_* + \frac{1}{2}\mathbf{t}'_*\boldsymbol{\Sigma}\mathbf{t}_*} \\ &= e^{\mathbf{t}'(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}) + \frac{1}{2}\mathbf{t}'\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\mathbf{t}}, \end{aligned}$$

which is the MGF of a $N(\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ distribution. \square

REMARK : Theorem 4.8 can be used to **standardize** a normally distributed random vector.

- ▶ Let \mathbf{Z} be a $N(\mathbf{0}, \mathbf{I})$ distributed $(n \times 1)$ random vector, that is a vector of n uncorrelated $N(0, 1)$ distributed random variables.
- ▶ Then the $(n \times 1)$ random vector \mathbf{Y} with a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ distribution can be represented in terms of \mathbf{Z} as

$$\mathbf{Y} = \boldsymbol{\mu} + \mathbf{AZ}, \quad \text{where } \mathbf{A} \text{ is selected such that}^3 \quad \mathbf{AA}' = \boldsymbol{\Sigma}.$$

This follows from Theorem 4.8, since

$$\mathbf{Y} \sim N(\mathbf{A}\mathbf{0} + \boldsymbol{\mu}, \mathbf{A}\mathbf{I}\mathbf{A}') = N(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

- ▶ Furthermore, the inversion of the function $\mathbf{Y} = \boldsymbol{\mu} + \mathbf{AZ}$ **standardizes the normally distributed vector \mathbf{Y}**

$$\mathbf{A}^{-1}(\mathbf{Y} - \boldsymbol{\mu}) = \mathbf{Z} \sim N(\mathbf{0}, \mathbf{I}). \quad \diamond$$

³ $\mathbf{AA}' = \boldsymbol{\Sigma}$ denotes the **Cholesky decomposition**, where the Cholesky factor \mathbf{A} is a lower triangular matrix.

A further important property of the multivariate normal distribution is that **marginal densities** obtained from a multivariate normal are normal densities as stated in the following theorem.

THEOREM 4.9 *Let \mathbf{Z} be an n -dimensional $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed random variable, where*

$$\mathbf{Z} = \left[\begin{array}{c} \mathbf{Z}_{(1)} \\ \mathbf{Z}_{(2)} \end{array} \right] = \left[\begin{array}{c} (m \times 1) \\ (n-m) \times 1 \end{array} \right], \quad \boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{(1)} \\ \boldsymbol{\mu}_{(2)} \end{array} \right] = \left[\begin{array}{c} (m \times 1) \\ (n-m) \times 1 \end{array} \right], \quad \text{and} \quad \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right] = \left[\begin{array}{c|c} (m \times m) & m \times (n-m) \\ \hline (n-m) \times m & (n-m) \times (n-m) \end{array} \right].$$

Then the marginal pdf of $\mathbf{Z}_{(1)}$ is $N(\boldsymbol{\mu}_{(1)}, \boldsymbol{\Sigma}_{11})$, and the marginal PDF of $\mathbf{Z}_{(2)}$ is $N(\boldsymbol{\mu}_{(2)}, \boldsymbol{\Sigma}_{22})$.

PROOF: Let

$$\mathbf{A} = \left[\begin{array}{c|c} \mathbf{I} & \mathbf{0} \\ \hline (m \times m) & m \times (n-m) \end{array} \right] \quad \text{and} \quad \mathbf{b} = \mathbf{0}$$

in Theorem 4.8, where \mathbf{I} is an identity matrix. It follows that $\mathbf{Z}_{(1)} = \mathbf{AZ}$ is $N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$ distributed, with $\mathbf{A}\boldsymbol{\mu} = \boldsymbol{\mu}_{(1)}$ and $\mathbf{A}\boldsymbol{\Sigma}\mathbf{A}' = \boldsymbol{\Sigma}_{11}$. (The proof for $\mathbf{Z}_{(2)}$ is analogous.) \square

REMARK : Note that Theorem 4.9 can be applied to obtain the marginal pdf of *any subset* of the normal random variable (Z_1, \dots, Z_n) by simply ordering them appropriately in the definition of Z in the theorem.

Also note that the normal marginal pdfs are derived from joint multivariate normality. The derivation does not go in the opposite direction. That is, marginal normality does not imply joint normality (for an example – see Casella and Berger, 2002, Exercise 4.47). ◇

The following theorem states an important result concerning *conditional densities* obtained from multivariate normal distributions.



THEOREM 4.10 Let \mathbf{Z} be an n -dimensional $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed random variable, where

$$\mathbf{Z} = \left[\begin{array}{c} \mathbf{Z}_{(1)} \\ \hline \mathbf{Z}_{(2)} \end{array} \right] \begin{array}{c} (m \times 1) \\ \\ (n-m) \times 1 \end{array}, \quad \boldsymbol{\mu} = \left[\begin{array}{c} \boldsymbol{\mu}_{(1)} \\ \hline \boldsymbol{\mu}_{(2)} \end{array} \right] \begin{array}{c} (m \times 1) \\ \\ (n-m) \times 1 \end{array}, \text{ and } \boldsymbol{\Sigma} = \left[\begin{array}{c|c} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \hline \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{array} \right] \begin{array}{c} (m \times m) \quad m \times (n-m) \\ \\ (n-m) \times m \quad (n-m) \times (n-m) \end{array};$$

and let \mathbf{z}^0 be an n -dimensional vector of constants partitioned conformably with the partition \mathbf{Z} into $\mathbf{z}_{(1)}^0$ and $\mathbf{z}_{(2)}^0$.

Then the conditional distributions of $\mathbf{Z}_{(1)}|\mathbf{Z}_{(2)} = \mathbf{z}_{(2)}^0$ and $\mathbf{Z}_{(2)}|\mathbf{Z}_{(1)} = \mathbf{z}_{(1)}^0$ are

$$\mathbf{Z}_{(1)}|(\mathbf{Z}_{(2)} = \mathbf{z}_{(2)}^0) \sim N\left(\boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\left[\mathbf{z}_{(2)}^0 - \boldsymbol{\mu}_{(2)}\right], \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\right)$$

$$\mathbf{Z}_{(2)}|(\mathbf{Z}_{(1)} = \mathbf{z}_{(1)}^0) \sim N\left(\boldsymbol{\mu}_{(2)} + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\left[\mathbf{z}_{(1)}^0 - \boldsymbol{\mu}_{(1)}\right], \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\right).$$

PROOF: By definition, the conditional distribution of $\mathbf{Z}_{(1)}|\mathbf{Z}_{(2)} = \mathbf{z}_{(2)}^0$ obtains as

$$\begin{aligned} f(\mathbf{z}_{(1)}|\mathbf{z}_{(2)} = \mathbf{z}_{(2)}^0) &= \frac{f(\mathbf{z}_{(1)}, \mathbf{z}_{(2)}^0)}{f(\mathbf{z}_{(2)}^0)} \\ &= \frac{\frac{1}{(2\pi)^n/2} |\boldsymbol{\Sigma}|^{1/2} \exp -\frac{1}{2} \left[\begin{matrix} \mathbf{z}_{(1)} - \boldsymbol{\mu}_{(1)} \\ \mathbf{z}_{(2)}^0 - \boldsymbol{\mu}_{(2)} \end{matrix} \right]' \boldsymbol{\Sigma}^{-1} \left[\begin{matrix} \mathbf{z}_{(1)} - \boldsymbol{\mu}_{(1)} \\ \mathbf{z}_{(2)}^0 - \boldsymbol{\mu}_{(2)} \end{matrix} \right]}{\frac{1}{(2\pi)^{(n-m)/2} |\boldsymbol{\Sigma}_{22}|^{1/2} \exp -\frac{1}{2} \left[\mathbf{z}_{(2)}^0 - \boldsymbol{\mu}_{(2)} \right]' \boldsymbol{\Sigma}_{22}^{-1} \left[\mathbf{z}_{(2)}^0 - \boldsymbol{\mu}_{(2)} \right]}}. \end{aligned}$$

The determinant $|\boldsymbol{\Sigma}|$ and the inverse $\boldsymbol{\Sigma}^{-1}$ can be *partitioned* as⁴

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{22}| \cdot |\boldsymbol{\Sigma}_{11 \cdot 2}|,$$

$$\boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11 \cdot 2}^{-1} & -\boldsymbol{\Sigma}_{11 \cdot 2}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \\ -\boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11 \cdot 2}^{-1} & \boldsymbol{\Sigma}_{22}^{-1} + \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21} \boldsymbol{\Sigma}_{11 \cdot 2}^{-1} \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \end{bmatrix},$$

where

$$\boldsymbol{\Sigma}_{11 \cdot 2} = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{21}.$$

Inserting the partitioned determinant and inverse and collecting terms produces the conditional density as stated in the theorem.

(CONTINUES)

⁴See Lütkepohl, H. (1996, p. 30 and 50), Handbook of Matrices, Chichester.

PROOF (CONTINUED): The proof for the conditional distribution of $\mathbf{Z}_{(2)}|\mathbf{Z}_{(1)} = \mathbf{z}_{(1)}^0$ is analogous. \square

REMARK : Note that the mean of the conditional distribution given by

$$E(\mathbf{Z}_{(1)}|\mathbf{z}_{(2)}) = \boldsymbol{\mu}_{(1)} + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1} \left(\mathbf{z}_{(2)} - \boldsymbol{\mu}_{(2)} \right)$$

is a **linear function** in the 'conditioning variable' $\mathbf{z}_{(2)}$.

This linearity of the conditional mean is a specific feature of the multivariate normal distribution as a member of the family of *elliptically contoured distributions*. \diamond

REMARK : Consider the special case where $Z_{(1)}$ is a scalar and $\mathbf{Z}_{(2)}$ is a $(k \times 1)$ vector.

Then the conditional mean of $Z_{(1)}$ given $\mathbf{z}_{(2)}$, that is, the regression function of $Z_{(1)}$ on $\mathbf{Z}_{(2)}$ has the form

$$E(Z_{(1)}|\mathbf{z}_{(2)}) = \underset{(1 \times 1)}{\mathbf{a}} + \underset{(1 \times k)}{\mathbf{b}} \mathbf{z}_{(2)},$$

where

$$\mathbf{a} = \mu_{(1)} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\mu}_{(2)}, \quad \mathbf{b} = \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}.$$

Hence, the regression function of $Z_{(1)}$ on $\mathbf{Z}_{(2)}$ obtained from a multivariate normal distribution for $(Z_{(1)}, \mathbf{Z}_{(2)})$ is linear. \diamond

The following theorem states that in the case of a normal distribution, zero covariance implies independence of the random variables, which in general is not true for other distributions.

THEOREM 4.11 Let $\mathbf{X} = (X_1, \dots, X_n)'$ be a $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ -distributed random variable. Then (X_1, \dots, X_n) are independent iff $\boldsymbol{\Sigma}$ is a diagonal matrix with all covariances being zero.

PROOF: To see that under normality zero covariances imply independence, consider the joint pdf

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}.$$

The diagonality of $\boldsymbol{\Sigma}$ implies that



$$|\boldsymbol{\Sigma}| = \prod_{i=1}^n \sigma_i^2, \quad \boldsymbol{\Sigma}^{-1} = \begin{pmatrix} 1/\sigma_1^2 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 1/\sigma_n^2 \end{pmatrix}.$$

Therefore, the joint pdf factors into the product

$$f(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^n \frac{1}{(2\pi)^{1/2} \sigma_i} \exp \left\{ -\frac{1}{2} \frac{(x_i - \mu_i)^2}{\sigma_i^2} \right\} = \prod_{i=1}^n N(\mu_i, \sigma_i^2),$$

which is the product of the n marginal densities. This implies independence.

(CONTINUES)

PROOF (CONTINUED): The proposition that **independence implies zero covariances** was proven in Chapter 3 (Theorem 3.20) and is true for any joint distribution. \square

The majority of families of distributions introduced so far are special cases of the **exponential class of distributions**.

As we shall see later, the families of distributions from the exponential class have many nice statistical properties which often simplify procedures for statistical inference (see Statistics IV).

DEFINITION (EXPONENTIAL CLASS OF PDFS): The pdf $f(\mathbf{x}; \Theta)$ is a member of the exponential class of pdfs iff it has the form

$$f(\mathbf{x}; \Theta) = \begin{cases} \exp \left\{ \sum_{i=1}^k c_i(\Theta) g_i(\mathbf{x}) + d(\Theta) + z(\mathbf{x}) \right\} & \text{for } \mathbf{x} \in A \\ 0 & \text{otherwise} \end{cases},$$

where

$$\mathbf{x} = (x_1, \dots, x_n)';$$

$$\Theta = (\Theta_1, \dots, \Theta_k)';$$

$c_i(\Theta), d(\Theta)$: real-valued functions of Θ that do not depend on \mathbf{x} ;

$g_i(\mathbf{x}), z(\mathbf{x})$: real-valued functions of \mathbf{x} that do not depend on Θ ;

$A \subset \mathbb{R}^n$: a range/support which does not depend on Θ .

REMARK : In order to check whether a family of pdfs belongs to the exponential class, we must identify the functions $c_i(\boldsymbol{\Theta})$, $d(\boldsymbol{\Theta})$, $g_i(\mathbf{x})$, $z(\mathbf{x})$ and show that the family has pdfs of the form given in the definition, which is not always trivial. ◇

EXAMPLE: Consider a **univariate $N(\mu, \sigma^2)$** -distribution with $n = 1$ and $k = 2$ (# of params). Set

$$\begin{aligned} c_1(\boldsymbol{\Theta}) &= \frac{\mu}{\sigma^2}, & c_2(\boldsymbol{\Theta}) &= -\frac{1}{2\sigma^2}, & g_1(x) &= x, & g_2(x) &= x^2; \\ d(\boldsymbol{\Theta}) &= -\frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{\mu^2}{\sigma^2}, & z(x) &= 0, & A &= \mathbb{R}^1. \end{aligned}$$

Substitution into the definition of the exponential class yields

$$\begin{aligned} f(x; \boldsymbol{\Theta}) &= \exp\left\{ \frac{\mu}{\sigma^2} \cdot x - \frac{1}{2\sigma^2} \cdot x^2 - \frac{1}{2} \ln(2\pi\sigma^2) - \frac{1}{2} \frac{\mu^2}{\sigma^2} \right\} \\ &= \frac{1}{(2\pi)^{1/2} \sigma} \exp\left\{ -\frac{1}{2} \frac{(x - \mu)^2}{\sigma^2} \right\}. \quad || \end{aligned}$$

REMARK : Further members of the exponential class are:

Bernoulli, binomial, multinomial, negative binomial, Poisson, geometric, gamma, chi-square, exponential, beta, etc.

Families of distributions that do not belong to the exponential class are, e.g.:

discrete uniform, continuous uniform, hypergeometric. \diamond