Introduction to Statistics

Change of Variables
Moments

LV Nr. 105.692
Summer Semester 2021

Efstathia Bura
E105
Institute of Statistics and Mathematical Methods in Economics

Outline

Transformations and Change of Variable

- Suppose X is a random variable with cdf $F_X(x)$ and another random variable Y defined by

$$Y = g(X),$$

where g is an arbitrary function.

  - Every function of random variables is a random variable!

- How to calculate probabilities for Y in terms of probabilities for X:

$$\mathbb{P}(Y \in A) = \mathbb{P}(g(X) \in A) \tag{1}$$

This contains the heart of the most general change of variable formula.

Change of Variable

- Let $\mathbb{P}_X$ denote the probability measure for the model in which X is the identity random variable,

$$X(\omega) = \omega$$

  and similarly $\mathbb{P}_Y$ for the analogous measure for Y.
- Then the left hand side of (1) is $\mathbb{P}_Y(A)$ and the right hand side is $\mathbb{P}_X(B)$, where

$$B = \{\omega \in \Omega : g(\omega) \in A\}$$

  where Ω is the sample space of the probability model describing X. Putting this all together, we get the following theorem.

Change of Variables: General Definition

Theorem

If $X \sim \mathbb{P}_X$ *and* $Y = g(X)$, *then* $Y \sim \mathbb{P}_Y$ *where*

$$\mathbb{P}_Y(A) = \mathbb{P}_X(B)$$

where the relation between A *and* B *is given by* $B = \{\omega \in \Omega : g(\omega) \in A\}$.

This theorem is too abstract for everyday use.
In practice, we will use other theorems that handle special cases more easily.
But we should keep in mind that this theorem exists and *allows*, at least in theory, the calculation of the distribution of *any random variable*.

Example: Constant Random Variable

Although the theorem is hard to apply to complicated random variables, it is not too hard for simple ones.

- The simplest random variable is a constant: $g(\omega) = c$, $\forall \omega \in \Omega$.
- To apply the theorem, we have to find, for any set A in the sample of Y, the set B.
- This sounds complicated, and in general it is, but here is it fairly easy: only two cases.

Example: Constant Random Variable

Case I: Suppose $c \in A$. Then,

$$B = \{\omega \in \Omega : g(\omega) \in A\} = \Omega$$

because $g(\omega) = c \in A$ for all $\omega \in \Omega$.

Case II: Suppose $c \notin A$. Then,

$$B = \{\omega \in \Omega : g(\omega) \in A\} = \varnothing$$

because $g(\omega) = c \notin A$ for all $\omega \in \Omega$; that is, there is no ω such that the condition holds, so the set of ω satisfying the condition is empty.

Combining the Cases: For any probability distribution the empty set has probability zero and the sample space has probability one. Thus the theorem says

$$\mathbb{P}_Y(A) = \begin{cases} 1, & c \in A \\ 0, & c \notin A \end{cases}$$

Even constant random variables have probability distributions. They are rather trivial, all the probabilities being either zero or one, but they are probability models that satisfy the axioms.

Discrete Random Variables

- For a discrete r.v. X,

$$\mathbb{P}(A) = \sum_{x \in A} f_X(x)$$

  where fX is the pmf of X ($f_X(x) = \mathbb{P}(X = x)$).
- Note also that for discrete probability models, not only is the above giving the measure in terms of the density, but also

$$f(x) = \mathbb{P}(\{x\})$$

  giving the mass in terms of the measure by taking $A = \{x\}$.
- We only need to consider sets A in the statement of the theorem that are one-point sets, which gives the following theorem.

Discrete Random Variables

Theorem

If X is a discrete random variable with pmf f_X and sample space S, and $Y = g(X)$, then Y is a discrete random variable with pmf f_Y defined by

$$f_Y(y) = \mathbb{P}_X(B) = \sum_{x \in B} f_X(x)$$

where $B = \{x \in S : y = g(x)\}$.

In other words,

$$f_Y(y) = \sum_{x \in S, y = g(x)} f_X(x)$$

Even with the simplification, this theorem is still a bit too abstract and complicated for general use.

One-To-One Transformations

- A transformation (change of variable) $g : S \to T$ is one-to-one if it maps each point x to a different value $g(x)$ from all other points, that is,

$$g(x_1) \neq g(x_2), \quad \text{for } x_1 \neq x_2$$

- **Example:** The function $g : \mathbb{R} \to \mathbb{R}$, $g(x) = x^2$ is **not** one-to-one because

$$g(x) = g(-x), \quad x \in \mathbb{R}$$

but $g : (0, \infty) \to \mathbb{R}$ defined by the very same formula **is** *one-to-one*!

Inverse Transformations

- A function is invertible if it is *one-to-one* and *onto*, the latter meaning that its codomain is the same as its range.
- Neither of the functions considered in the previous example are invertible. The second is one-to-one, but it is not onto, because g maps positive real numbers to positive real numbers.
- To obtain a function that is invertible, we need to restrict the codomain to be the same as the range, defining the function

$$g : (0, \infty) \to (0, \infty)$$

$$g(x) = x^2$$

Inverse Transformations

- Every invertible function

$$g : S \to T$$

has an inverse function

$$g^{-1} : T \to S$$

satisfying

$$g(g^{-1}(y)) = y, \quad y \in T$$

and

$$g^{-1}(g(x)) = x, \quad x \in S$$

- To obtain the inverse we solve

$$y = g(x)$$

for x. For example, if $y = g(x) = x^2$, then

$$x = \sqrt{y} = g^{-1}(y)$$

- If for any y there is no solution or multiple solutions, the inverse does not exist (if no solutions the function is not onto, if multiple solutions it is not one-to-one).

Change of Variable for Invertible Transformations

- For invertible transformations the change of variable theorem simplifies considerably.
- The set B in the theorem is always a singleton: there is a unique x such that $y = g(x)$, namely $g^{-1}(y)$. So

$$B = \{g^{-1}(y)\}$$

and the theorem can be stated as follows.

Theorem

If X is a discrete random variable with pmf f_X and sample space Ω, if $g : \Omega \to T$ is an invertible transformation and $Y = g(X)$, then Y is a discrete random variable with pmf f_Y defined by

$$f_Y(y) = f_X(g^{-1}(y)), \quad y \in T$$

Example: the "other" geometric

- Suppose X has pmf

$$f_X(x) = (1-p)p^x, \quad x = 0, 1, \ldots$$

  for $p \in (0, 1)$.
- Some people like to start counting at one rather than zero and prefer to call the distribution of the random variable $Y = X + 1$ the "geometric distribution"
- The transformation in question is quite simple

$$y = g(x) = x + 1$$

  with inverse

$$x = g^{-1}(y) = y - 1$$

$$g : \{0, 1, \ldots\} \to \{1, \ldots\}.$$

Example: the "other" geometric

Now we just apply the theorem.

$$f_Y(y) = f_X(g^{-1}(y)) = (1-p)p^{y-1}, \quad y-1 = 0, 1, \ldots$$

or,

$$f_Y(y) = (1-p)p^{y-1}, \quad y = 1, 2, \ldots$$

Continuous Random Variables

For continuous random variables, probability measures are defined by integrals

$$\mathbb{P}(A) = \int_A f_X(x)dx$$

where f_X is the pdf of X.
The analogous theorem is

Theorem

Suppose X is a continuous random variable with pmf f_X and sample space S. If $g : S \to T$ is an invertible transformation with differentiable inverse $h = g^{-1}$, and $Y = g(X)$, then Y is a continuous random variable with pdf f_Y defined by

$$f_Y(y) = f_X(g^{-1}(y))|(g^{-1}(y))'| = f_X(h(y))|h'(y)|, \quad y \in T$$

Example: Exponential

Suppose $X \sim Exp(\lambda)$ with $f_X(x) = \lambda e^{-\lambda x}$. What is the distribution of $Y = X^2$?

- The transformation in question is $g : (0, \infty) \to (0, \infty)$ defined by
$$g(x) = x^2, \quad x > 0$$
- The inverse transformation is
$$h(y) = g^{-1}(y) = y^{1/2}, \quad y > 0$$
- Its derivative is
$$h'(y) = \frac{1}{2} y^{-1/2}, \quad y > 0$$
- Applying the change-of-variables theorem with $h(y) = \sqrt{y}$ gives
$$\begin{aligned} f_Y(y) &= f_X(h(y))|h'(y)| \\ &= \lambda e^{-\lambda\sqrt{y}} \frac{1}{2} y^{-1/2} \\ &= \frac{\lambda e^{-\lambda\sqrt{y}}}{2\sqrt{y}}, \quad y > 0 \end{aligned}$$

Random Vectors

- A vector is a mathematical object consisting of a sequence of real numbers. We usually write vectors using boldface type

$$\mathbf{x} = (x_1, \ldots, x_n)^T$$

- The separate numbers $x_1, \ldots, x_n$ are called the components or coordinates of the vector.
- A random vector is simply a vector-valued random variable.
  - We denote random vectors by capital letters and their possible values by lower case letters. So a random vector

$$\mathbf{X} = (X_1, \ldots, X_n)^T$$

    is a vector whose components are real-valued random variables $X_1, \ldots, X_n$.

Discrete Random Vectors

A real-valued function f on a countable subset S of $\mathbb{R}^n$ is the probability mass function of a discrete random vector if it satisfies the following two properties

$$f(\mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \in S$$
$$\sum_{\mathbf{x} \in S} f(\mathbf{x}) = 1$$

The corresponding probability measure is defined by

$$\mathbb{P}(A) = \sum_{\mathbf{x} \in A} f(\mathbf{x})$$

for all events A (subsets of the sample space S).

Continuous Random Vectors

Similarly, a real-valued function f on a subset S of $\mathbb{R}^n$ is the probability density function of a continuous random vector if it satisfies the following two properties:

$$f(\mathbf{x}) \geq 0, \quad \text{for all } \mathbf{x} \in S$$

$$\int_{\mathbf{x} \in S} f(\mathbf{x}) = 1$$

The corresponding probability measure is defined by

$$\mathbb{P}(A) = \int_A f(\mathbf{x}) d\mathbf{x} = \int \int_A \ldots \int f(x_1, \ldots, x_n) dx_1 dx_2 \ldots dx_n$$

for all events A (subsets of the sample space S).

Example

Suppose that f is the probability density on the unit square in $\mathbb{R}^2$ defined by

$$f(x, y) = x + y, \quad 0 < x < 1,\ 0 < y < 1$$

Calculate $\mathbb{P}(X + Y > 1)$.

- Let

$$A = \{(x, y) : 0 < x < 1, 0 < y < 1 \text{ and } x + y > 1\}$$

- Then, $\mathbb{P}(X + Y > 1) = \mathbb{P}(A)$
- We keep x fixed, and consider y variable. What are the limits of the values of y?

$$0 < y < 1,\ 1 < x + y \Rightarrow 1 - x < y < 1$$

Example

$$\mathbb{P}(A) = \int_0^1 \int_{1-x}^1 f(x,y)dydx$$

The inner integral is

$$\int_{1-x}^1 (x+y)dy = xy + \frac{y^2}{2}\Big|_{1-x}^1 = (x+2) - \left(x(1-x) + \frac{(1-x)^2}{2}\right)$$
$$= x + \frac{x^2}{2}$$

So,

$$\int_0^1 \left(x + \frac{x^2}{2}\right) dx = \frac{x^2}{2} + \frac{x^3}{6}\Big|_0^1 = \frac{2}{3}$$

- In more complicated situations, finding the limits of integration can be much trickier, but this is rare in probability and statistics.

The Support of a Random Variable

The support of a random variable is the set of points where its density (pmf) is positive.
If a random variable X has support A, then $\mathbb{P}(X \in A) = 1$, because if S is the sample space for the distribution of X

$$\begin{aligned} 1 &= \int_S f_X(x)dx \\ &= \int_A f_X(x)dx + \int_{A^c} f_X(x)dx \\ &= \int_A f_X(x)dx \\ &= \mathbb{P}(X \in A) \end{aligned}$$

Thus, as long as the only random variables under consideration are X and functions of X it makes no difference whether we consider the sample space to be S (the original sample space) or A (the support of X). We can use this observation in two ways.

1. If the support of a random variable is not the whole sample space, we can throw the points where the density is zero out of the sample space without changing any probabilities.
2. Conversely, we can always consider a random variable to live in a larger sample space by defining the density to be zero outside of the original sample space.

Example: Uniform Distribution

$$X \sim U(a, b)$$

1. We can consider the sample space to be the interval (a, b), in which case we write the density

$$f(x) = \frac{1}{b-a}, \quad a < x < b$$

2. we may want to consider the sample space to be the whole real line, in which case we can write the density in two different ways, one using case splitting

$$f(x) = \begin{cases} 0, & x \leq a \\ \frac{1}{b-a}, & a < x < b \\ 0, & b \leq x \end{cases}$$

3. and the other using indicator functions

$$f(x) = \frac{1}{b-a} I_{(a,b)}(x)$$

Joint and Marginal Distributions

1. When two different sets are under discussion, one a subset of the other, we use joint to indicate the superset and marginal to indicate the subset.
2. For example, if we are interested in the distribution of the random variables X, Y, and Z and simultaneously interested in the distribution of X and Y, then we call the distribution of the three variables with density $f_{X,Y,Z}$ the joint distribution and density, whereas we call the distribution of the two variables X and Y with density $f_{X,Y}$ the marginal distribution and density.
    1. What is the relationship between joint and marginal densities? Given $f_{X,Y}$ how do we obtain f_X?
    2. this is a question about change of variables:

$$X = g(X, Y)$$

Joint and Marginal Distributions

1. The change of variables theorem applied to this case says that

$$\mathbb{P}_X(A) = \mathbb{P}_{X,Y}(B)$$

where

$$\begin{aligned} B &= \{(x,y) \in \mathbb{R}^2 : g(x,y) \in A\} \\ &= \{(x,y) \in \mathbb{R}^2 : x \in A\} \\ &= A \times \mathbb{R} \end{aligned}$$

2. Now the definition of the density of a continuous random variable gives us

$$\mathbb{P}_X(A) = \int_A f_X(x)dx$$

3. whereas the definition of the density of a continuous (bivariate) random vector gives

$$\begin{aligned} \mathbb{P}_{X,Y}(B) &= \int\int_B f_{X,Y}(x,y)dxdy = \int\int_{A\times\mathbb{R}} f_{X,Y}(x,y)dxdy \\ &= \int_A \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx \end{aligned}$$

Joint and Marginal Distributions

1. Thus we can calculate $\mathbb{P}(X \in A)$ in two diffeerent ways, which must be equal

$$\int_A f_X(x)dx = \int_A \int_{-\infty}^{\infty} f_{X,Y}(x,y)dydx$$

2. Equality of the two expressions for arbitrary events A requires that $f_X(x)$ be the result of the y integral,

$$f_X(x) = \int f_{X,Y}(x,y)dy$$

3. That is, *to go from joint to marginal you integrate (or sum) out the variables you don't want:*

$$\int f_{X,Y}(x,y)dy = \text{ some function of x only}$$

4. We sum out discrete variables and integrate out continuous ones.

Uniform Distribution on a Triangle

Consider the uniform distribution on the triangle with corners (0, 0), (1, 0), and (0,1), and density

$$f(x,y) = 2, \quad 0 < x, y \text{ and } x + y < 1$$

What is the marginal distribution of X?

1. We have to figure out the limits of integration.
2. Clearly $x > 0$ is required. Also we must have $x < 1 - y$. This inequality is least restrictive when we take $y = 0$. So the range of the random variable X is $0 < x < 1$. Then, $0 < y < 1 - x$:

$$\begin{aligned} f_X(x) &= \int_0^{1-x} f_{X,Y}(x,y)dy = \int_0^{1-x} 2dy \\ &= 2y\Big|_0^{1-x} = 2(1-x), \quad 0 < x < 1 \end{aligned}$$

3. The marginal is not uniform, although the joint is uniform!

Derivatives of Vector Functions

1. Let $\mathbf{g}$ be a function that maps n-dimensional vectors to m-dimensional vectors: $\mathbf{g} : \mathbb{R}^n \to \mathbb{R}^m$

$$\mathbf{g}(\mathbf{x}) = \begin{pmatrix} g_1(\mathbf{x}) \\ g_2(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{pmatrix}$$

2. The derivative of the function $\mathbf{g}$ at the point $\mathbf{x}$ (assuming it exists) is the matrix of partial derivatives:

$$\mathbf{G} = \nabla \mathbf{g}(\mathbf{x}) = (g_{ij}) = \left(\frac{\partial g_i(\mathbf{x})}{\partial x_j} \right) : m \times n$$

Invertible Transformations

1. A multivariate change of variables $\mathbf{h}$ cannot be invertible unless it maps between spaces of the same dimension:

$$\mathbf{h} : \mathbb{R}^n \to \mathbb{R}^n, \quad \text{for some } n.$$

2. The determinant of its derivative matrix is called the Jacobian of the mapping, denoted

$$J(\mathbf{x}) = \det\left(\nabla \mathbf{h}(\mathbf{x})\right)$$

Change of Variable in Integration

Theorem

Suppose that $\mathbf{h}$ *is an invertible, continuously differentiable mapping with nonzero Jacobian defined on an open subset of* $\mathbb{R}^n$*, and suppose that* A *is a region contained in the domain of* $\mathbf{h}$ *and that* f *is an integrable function defined on* $\mathbf{h}(A)$*. Then,*

$$\int_{\mathbf{h}(A)} f(\mathbf{x})d\mathbf{x} = \int_A f(\mathbf{h}(\mathbf{y}))\,|J(\mathbf{y})|d\mathbf{y}$$

where J *is the Jacobian of* $\mathbf{h}$*.*

Change of Variable for Densities

Theorem

Suppose that $\mathbf{g}$ *is an invertible mapping defined on an open subset of* $\mathbb{R}^n$ *containing the support of a continuous random vector* $\mathbf{X}$ *with probability density* $f_{\mathbf{X}}$*, and suppose that* $\mathbf{h} = \mathbf{g}^{-1}$ *is continuously differentiable with nonzero Jacobian* J*. Then, the random vector* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *has probability density*

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{\mathbf{X}}(\mathbf{h}(\mathbf{y}))\,|J(\mathbf{y})| = f_{\mathbf{X}}(\mathbf{h}(\mathbf{y}))\,|det(\nabla \mathbf{h}(\mathbf{y}))|$$

The univariate change-of-variable formula is a special case.

Proof

The general change of variable theorem obtains

$$\mathbb{P}_{\mathbf{Y}}(A) = \mathbb{P}_{\mathbf{X}}(B)$$

where

$$B = \{\mathbf{x} \in S : \mathbf{g}(\mathbf{x}) \in A\}$$

where S is the sample space of the random vector $\mathbf{X}$, which we may take to be the open subset of $\mathbb{R}^n$ on which g is defined. Because $\mathbf{g}$ is invertible,

$$B = \mathbf{h}(A), \quad A = \mathbf{g}(B)$$

Using the definition of measures in terms of densities gives

$$\int_A f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = \int_B f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} = \int_{\mathbf{h}(A)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$$

Applying the *Change of Variable in Integration* theorem to the right hand side,

$$\int_A f_{\mathbf{Y}}(\mathbf{y}) d\mathbf{y} = \int_A f_{\mathbf{X}}(\mathbf{h}(\mathbf{y})) \, |J(\mathbf{y})|$$

This is true for all sets A only if the integrands are equal, which is the assertion of the theorem.

Example

Suppose

$$f(x,y) = \frac{1}{2\pi} \exp\left(-\frac{x^2}{2} - \frac{y^2}{2}\right), \quad (x,y) \in \mathbb{R}^2$$

Find the joint density of the variables

$$U = X, \quad V = Y/X$$

This transformation is undefined when $X = 0$, but that event occurs with probability zero and can be ignored.
The inverse transformation is

$$X = U, \quad Y = UV$$

with derivative

$$\begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ v & u \end{pmatrix}$$

and Jacobian $1 \cdot u - v \cdot 0 = u$.
The joint density of U and V is

$$\begin{aligned} g(u,v) &= \frac{1}{2\pi} \exp\left(-\frac{u^2}{2} - \frac{(uv)^2}{2}\right) |u| \\ &= \frac{|u|}{2\pi} \exp\left(-\frac{u^2(1+v^2)}{2}\right) \end{aligned}$$

Example: convolution formula

Theorem

If X and Y are independent continuous real-valued random variables with densities f_X and f_Y, then $X+Y$ has density

$$f_{X+Y}(z) = \int f_X(z-y) f_Y(y) dy$$

This is called the convolution formula, and the function f_{X+Y} is called the convolution of the functions f_X and f_Y.

Proof: Let

$$u = x + y$$
$$v = y$$

with inverse mapping

$$x = u - v$$
$$y = v$$

The Jacobian is

$$J(u,v) = \left| \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} \right| = \left| \begin{pmatrix} 1 & -1 \\ 0 & 1 \end{pmatrix} \right| = 1$$

By the change-of-variable formula, the joint density of U and V is

$$f_{U,V}(u,v) = f_{X,Y}(u-v,u)|J(u,v)| = f_X(u-v)f_Y(v)$$

by the independence of X and Y. We find the marginal of U by integrating out V

$$f_U(u) = \int f_X(u-v)f_Y(v)dv$$

which is the convolution formula. □

The discrete convolution formula is:

$$\mathbb{P}(U=z) = \sum_y f_Y(y)f_X(z-y)$$

Noninvertible Transformations

When $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ is not invertible, things are much more complicated, except in one special case, which is covered next:

$$g : \mathbb{R} \to [0, \infty)$$

with

$$g(x) = x^2, \quad x \in \mathbb{R}$$

This function is not invertible, because it is not one-to-one, but it has two *right inverses*, defined by

$$h_+(y) = \sqrt{y}, \quad y \geq 0$$

and

$$h_-(y) = -\sqrt{y}, \quad y \geq 0$$

Right inverses because

$$g(h_+(y)) = g(h_-(y)) = y$$

A partition of a set S is a family of sets $\{A_i : i \in I\}$ that are disjoint and cover S, that is,

$$A_i \cap A_j = \varnothing, \quad i \in I, j \in I, i \neq j$$

and

$$\cup_{i \in I} A_i = S$$

Theorem

Suppose $\mathbf{g} : U \to V$ *is a mapping, where* U *and* V *are open subsets of* $\mathbb{R}^n$*, and* U *is the support of a continuous random vector* $\mathbf{X}$*, i.e.* $\mathbb{P}(X \in A) = 1$*, with probability density* $f_{\mathbf{X}}$*. Suppose that* $\mathbf{h}_i, i \in I$ *are continuously differentiable right inverses of* $\mathbf{g}$ *with nonzero Jacobians* $J_i = det(\nabla \mathbf{h}_i)$*, and suppose the sets* $\mathbf{h}_i(V), i \in I$ *form a partition of* U*. Then, the random vector* $\mathbf{Y} = \mathbf{g}(\mathbf{X})$ *has pdf*

$$f_{\mathbf{Y}}(\mathbf{y}) = \sum_{i \in I} f_{\mathbf{X}}(\mathbf{h}_i(\mathbf{y})) \, |J_i(\mathbf{y})|$$

Proof.

The proof starts just like the previous proofs. We still have

$$\mathbb{P}_{\mathbf{Y}}(A) = \mathbb{P}_{\mathbf{X}}(B)$$

where

$$B = \{\mathbf{x} \in U : \mathbf{g}(\mathbf{x}) \in A\}$$

Now $\mathbf{g}$ is not invertible, but the sets $\mathbf{h}_i(A)$ form a partition of B. Hence,

$$\begin{aligned}
\int_A f_{\mathbf{Y}}(\mathbf{y}) &= \int_B f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \sum_{i \in I} \int_{\mathbf{h}_i(A)} f_{\mathbf{X}}(\mathbf{x}) d\mathbf{x} \\
&= \sum_{i \in I} \int_A f_{\mathbf{X}}(\mathbf{h}_i(\mathbf{y})) |J_i(\mathbf{y})| d\mathbf{y}
\end{aligned}$$

This is true for all sets A only if the integrands are equal, which is the assertion of the theorem. □

Example

Suppose X is a random variable with density

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}$$

What is the density of $Y = X^2$?

- In order to apply the theorem, we need to delete the point zero from the sample space of X, then the transformation

$$g : (-\infty, 0) \cup (0, \infty) \to (0, \infty)$$

defined by $g(x) = x^2$ has the two right inverses:

$$\begin{aligned} h_- : (-\infty, 0) \to (0, \infty) \quad & h_-(y) = -\sqrt{y} \\ h_+ : (0, \infty) \to (0, \infty) \quad & h_+(y) = \sqrt{y} \end{aligned}$$

Example

- The ranges of the right inverses form a partition of the domain of g and they have derivatives

$$h'_-(y) = -\frac{1}{2}y^{-1/2}$$
$$h'_+(y) = \frac{1}{2}y^{-1/2}$$

Hence,

$$f_Y(y) = f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}} = \frac{1}{\sqrt{y}}f_X(\sqrt{y})$$
$$= \frac{1}{\sqrt{2\pi y}}e^{-y/2}, \quad y > 0$$

because f_X is symmetric about zero, that is, $f_X(x) = f_X(-x)$.

Corollary

Suppose X is a continuous random scalar with density f_X, then $Y = X^2$ has density

$$f_Y(y) = \frac{1}{2\sqrt{y}} \left(f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right), \quad y > 0$$

Moreover, if f_X is symmetric about zero, then

$$f_Y(y) = \frac{1}{\sqrt{y}} f_X(\sqrt{y}), \quad y > 0$$

Change of Variable: 2nd Method

We can also use cdf's to calculate change-of-variable.

1. If X has pdf f_X and $Y = g(X)$, then the cdf of Y is

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$$

2. Then we can find the pdf of Y by differentiation.
3. Back to our example: Suppose $Y = X^2$ and X has pdf f_X. What is the pdf of Y?
4. Since g is not invertible if X takes both positive and negative values, the "Jacobian method" is not usable. We use the 2nd method.

Change of Variable: 2nd Method

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(X^2 \le y) \\
&= \mathbb{P}(-\sqrt{y} \le X \le \sqrt{y}) \\
&= F_X(\sqrt{y}) - F_X(-\sqrt{y}) \\
f_Y(y) &= \frac{dF_Y(y)}{dy} \\
&= \frac{d}{dy}\left(F_X(\sqrt{y}) - F_X(-\sqrt{y})\right) \\
&= f_X(\sqrt{y})\frac{1}{2\sqrt{y}} + f_X(-\sqrt{y})\frac{1}{2\sqrt{y}}
\end{aligned}
$$

If X is symmetric about zero, then

$$
f_Y(y) = f_X(\sqrt{y})\frac{1}{\sqrt{y}}, \quad y > 0
$$

The Chi-Square Distribution

If X is standard normal,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad -\infty < x < \infty$$

then the distribution of $Y = X^2$ is called the chi-squared distribution for one degree of freedom. It has pdf

$$\begin{aligned} f_Y(y) &= f_X(\sqrt{y}) \frac{1}{\sqrt{y}} \\ &= \frac{1}{\sqrt{2\pi}} y^{-1/2} e^{-y/2}, \quad y > 0 \end{aligned}$$

From the form of the pdf we see that this is another name for the Gamma$(1/2, 1/2)$ distribution.

The Gamma Distribution

The function

$$f(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x > 0$$

is the pdf of a random variable with the $Gamma(\alpha, \lambda)$ distribution with shape parameter α and rate parameter λ.
To show it is a proper density, let $y = \lambda x$,

$$\begin{aligned}
\frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-\lambda x} dx &= \frac{\lambda^\alpha}{\Gamma(\alpha)} \int_0^\infty \left(\frac{y}{\lambda}\right)^{\alpha-1} e^{-y} \frac{dy}{\lambda} \\
&= \frac{\lambda^\alpha}{\Gamma(\alpha)} \frac{\Gamma(\alpha)}{\lambda^\alpha} \\
&= 1
\end{aligned}$$

HW: Suppose X and Y are independent gamma random variables with $X \sim Gamma(\alpha_1, \lambda)$, $Y \sim Gamma(\alpha_2, \lambda)$. Show that $U = X + Y$, $V = X/(X + Y)$ are independent random variables with $U \sim Gamma(\alpha_1 + \alpha_2, \lambda)$.

Standard Normal Distribution

Proof that the standard normal pdf integrates to one:
Let

$$c = \int_{-\infty}^{\infty} e^{-x^2/2} dx$$

then

$$\begin{aligned} c^2 &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-x^2/2 - y^2/2} dx dy \\ &= \int_0^{2\pi} \int_0^{\infty} e^{-r^2/2} r dr d\theta \\ &= 2\pi \int_0^{\infty} e^{-r^2/2} r dr \\ &= 2\pi \left[-e^{-r^2/2} \right]_0^{\infty} \\ &= 2\pi \end{aligned}$$

Linear Change of Variable

Suppose X is a continuous random variable with pdf f_X which is defined on the whole real line. Then $Y = \mu + \sigma X$ has pdf

$$f_Y(y) = \frac{1}{|\sigma|} f_X\left(\frac{y-\mu}{\sigma}\right)$$

if $\sigma \neq 0$. (otherwise Y is the constant random variable always having the value μ and does not have a pdf).
Proof: $y = \mu + \sigma x \Rightarrow$

$$h(y) = \frac{y-\mu}{\sigma}$$

with Jacobian $1/\sigma$. We next apply the change-of-variable formula. As x goes from $-\infty$ to ∞ so does y, and vice versa. Hence the range of Y is the whole real line.

Location-Scale Families

The parametric family of distributions having pdf of the form

$$f_{\mu,\sigma}(y) = \frac{1}{\sigma} f\left(\frac{y-\mu}{\sigma}\right)$$

where

- μ and σ are parameters,
- μ is called the location and σ the scale parameter, with $\sigma > 0$, $\mu \in \mathbb{R}$,

is called the location-scale family with standard distribution having pdf $f = f_{0,1}$, which must be a pdf defined on the whole real line.

Location-Scale Families

A location-scale family we already know is the $U(a, b)$ family.
However, a and b are not location-scale pair of parameters.
We can take a to be the location parameter and $\sigma = b - a$ to be the scale parameter.
Then the standard continuous uniform distribution has $a = 0$ and $b - a = 1$, so $b = 1$, that is, the $U(0, 1)$ distribution is the standard one.

$$\begin{aligned} f_{a,\sigma}(x) &= \frac{1}{\sigma} f_{0,1}\left(\frac{x-a}{\sigma}\right) \\ &= \frac{1}{b-a} I_{(0,1)}\left(\frac{x-a}{b-a}\right) \\ &= \frac{1}{b-a} I_{(a,b)}(x) \end{aligned}$$

General Normal Distributions

The location-scale family whose standard pdf is the standard normal pdf, is called the family of normal distributions.
The normal distribution with location parameter μ and scale parameter σ is abbreviated $N(\mu, \sigma^2)$. It has pdf

$$\begin{aligned} f_{\mu,\sigma}(x) &= \frac{1}{\sigma} f_{0,1}\left(\frac{x-\mu}{\sigma}\right) \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty \end{aligned}$$

Symmetry

- We say a random variable X is symmetric about zero if -X has the same distribution as X.
  - If the distribution of X is specifed by a pmf or a pdf f, then the distribution is symmetric about zero if

$$f(x) = f(-x), \quad x \in S$$

- We say a random variable X is symmetric about the point a if $X - a$ is symmetric about zero, that is, if $-(X - a)$ has the same distribution as $X - a$.
  - In this case we say a is the center of symmetry of the distribution of X.

Symmetry

Some symmetric distributions:

- The discrete uniform distribution on $\{1, \ldots, n\}$ is symmetric about $(n+1)/2$
- The $Bin(n, p)$ distribution is symmetric about $n/2$ if $p = 1/2$
- $U(a, b)$ is symmetric about $(a+b)/2$
- The $N(\mu, \sigma^2)$ distribution is symmetric about μ
- The $Beta(\alpha_1, \alpha_2)$ distribution is symmetric about $1/2$ if $\alpha_1 = \alpha_2$

Expectation

A common goal is to understand or summarize the behaviour of a random variable. One way to do this is by trying to understand some type of "typical behaviour" of a random variable.
Let X be a random variable with cdf $F_X(x)$. The expectation or mean, or average, or first moment of a random variable $g(X)$ is defined as:

$$E(g(X)) = \begin{cases} \int_{-\infty}^{+\infty} g(x)\, f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_{x \in \mathcal{X}} g(x)\, f_X(x) & \text{if } X \text{ is discrete,} \end{cases}$$

provided that the integral and the sum exist.
The expectation does not exist if $\mathbb{E}|g(X)| = \infty$.

Moments

1. $m_k = \mathbb{E}(X^k)$, $k \in \mathbb{N}$ ... *moments of order* k
   for $k = 1$ we obtain the expectation of X, i.e.

$$m_1 = \mathbb{E}(X)$$

2. $\mu_k = \mathbb{E}(X - E(X))^k$, $k \in \mathbb{N}$ ... *central moments of order* k
   for $k = 2$ we obtain the *variance* of X, i.e.

$$\mu_2 = \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{V}ar(X).$$

   The positive square root of $\mathbb{V}ar(X)$ is the *standard deviation* of X.

Poisson

Let $X \sim \mathcal{P}(\lambda)$, i.e. the pmf of X is of the form

$$p_X(k) = P(X = k) = \frac{\lambda^k}{k!} \cdot e^{-\lambda} \quad \text{for} \quad k \in \mathbb{N}_0. \tag{2}$$

Then,

$$\begin{aligned}
\mathbb{E}X &= \sum_{k=0}^{\infty} k \cdot p_X(k) = e^{-\lambda} \cdot \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} = e^{-\lambda} \cdot \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} \\
&= e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^{k+1}}{k!} = \lambda e^{-\lambda} \cdot \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = \lambda e^{-\lambda} \cdot e^{\lambda} = \lambda \qquad \text{and} \\
\mathbb{V}ar(X) &= \mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X - \lambda)^2 = e^{-\lambda} \cdot \sum_{k=1}^{\infty} (k-\lambda)^2 \frac{\lambda^k}{k!} \\
&= e^{-\lambda} \cdot \left(\sum_{k=1}^{\infty} k^2 \frac{\lambda^k}{k!} - 2\lambda \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k}{k!} + \lambda^2 \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \right) \\
&= e^{-\lambda} \cdot \left(\sum_{k=1}^{\infty} k \frac{\lambda^k}{(k-1)!} - 2\lambda \mathbb{E}X \cdot e^{\lambda} + \lambda^2 \cdot e^{\lambda} \right) \\
&= e^{-\lambda} \cdot \left(\sum_{k=2}^{\infty} \frac{\lambda^k}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^k}{(k-1)!} - \lambda^2 \cdot e^{\lambda} \right) \\
&= e^{-\lambda} \cdot \left((\lambda^2 + \lambda) \cdot e^{\lambda} - \lambda^2 \cdot e^{\lambda} \right) = \lambda.
\end{aligned}$$

In both calculations, we used $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$.

Gamma

Let $X \sim Gamma(\alpha, \beta)$ with pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, x^{\alpha-1}\, e^{-\frac{x}{\beta}}, \quad x > 0,\ \alpha, \beta > 0, \tag{3}$$

where $\Gamma(\alpha)$ is the Euler Gamma function $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx$.
Then,

$$\begin{aligned}
\mathbb{E}X &= \int_{-\infty}^{+\infty} x f_X(x)\, dx = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \int_0^{+\infty} x^{\alpha}\, e^{-\frac{x}{\beta}}\, dx \\
&= [\text{substitution: } x = \beta u] = \frac{\beta}{\Gamma(\alpha)} \int_0^{+\infty} u^{\alpha}\, e^{-u}\, du \\
&= \frac{\beta}{\Gamma(\alpha)} \left(-u^{\alpha} e^{-u} \Big|_0^{+\infty} + \alpha \int_0^{+\infty} u^{\alpha-1}\, e^{-u}\, du \right) = \frac{\beta}{\Gamma(\alpha)} \cdot \alpha\, \Gamma(\alpha) = \alpha\beta.
\end{aligned}$$

Providing similar calculations (two partial integrations) we obtain

$$\mathbb{E}(X^2) = \int_{-\infty}^{+\infty} x^2 f_X(x)\, dx = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \int_0^{+\infty} x^{\alpha+1}\, e^{-\frac{x}{\beta}}\, dx = \alpha(\alpha+1)\beta^2$$

and thus

$$\mathbb{V}ar(X) = \mathbb{E}(X - \alpha\beta)^2 = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \int_0^{+\infty} (x - \alpha\beta)^2\, x^{\alpha-1}\, e^{-\frac{x}{\beta}}\, dx = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha^2\beta^2.$$

HW

1. [**HW**] Let X be a Cauchy random variable with pdf

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

Show that the expectation of X does not exist. What can be said for its higher moments?

2. [**HW**] The Pareto distribution with parameters α and β has pdf

$$f(x) = \frac{\beta\,\alpha^{\beta}}{x^{\beta+1}}, \quad \alpha < x < +\infty, \;\; \alpha > 0, \;\; \beta > 0.$$

    1. Verify that $f(x)$ is a pdf.
    2. Derive mean and variance of this distribution. Prove that the variance does not exist if $\beta \leq 2$.

Properties of Expectation

Theorem

Let X be a random variable and let a, b and c be real constants. Then for any functions $g_1(x)$ and $g_2(x)$ whose expectations exist the following properties hold:

(a) $E\left(ag_1(x) + bg_2(x) + c\right) = a\,Eg_1(x) + b\,Eg_2(x) + c$

(b) *If* $g_1(x) \geq 0$ *for all* x *then* $Eg_1(x) \geq 0$.

(c) *If* $g_1(x) \geq g_2(x)$ *for all* x *then* $Eg_1(x) \geq Eg_2(x)$.

(d) *If* $a \leq g_1(x) \leq b$ *for all* x *then* $a \leq Eg_1(x) \leq b$.

All derive from properties of integration.

Minimizing distance (MSE)

In this example we provide a very useful method of finding a *good predictor* of X, i.e. we show

$$\min_b \mathbb{E}(X-b)^2 = \mathbb{E}(X-\mathbb{E}X)^2. \tag{4}$$

$$\begin{aligned}
\mathbb{E}(X-b)^2 &= \mathbb{E}(X-\mathbb{E}(X)+\mathbb{E}(X)-b)^2 \\
&= \mathbb{E}\left((X-\mathbb{E}(X))+(\mathbb{E}(X)-b)\right)^2 \\
&= \mathbb{E}(X-\mathbb{E}(X))^2 + 2\mathbb{E}\left((X-\mathbb{E}(X))(\mathbb{E}(X)-b)\right) + \mathbb{E}(\mathbb{E}(X)-b)^2 \\
&= \mathbb{V}ar(X) + 2(\mathbb{E}(X)-b)\,\mathbb{E}(X-\mathbb{E}(X)) + (\mathbb{E}(X)-b)^2 \\
&= \mathbb{V}ar(X) + (\mathbb{E}(X)-b)^2,
\end{aligned}$$

since $(\mathbb{E}(X)-b)$ is constant and $\mathbb{E}(X-\mathbb{E}(X)) = \mathbb{E}(X)-\mathbb{E}(X) = 0$. Thus, we expressed $\mathbb{E}(X-b)^2$ in the form

$$\mathbb{E}(X-b)^2 = \mathbb{V}ar(X) + (\mathbb{E}(X)-b)^2. \tag{5}$$

The first term in (5) cannot be controlled as it does not depend on X, and the second term is nonnegative. Therefore, $\mathbb{E}(X-b)^2$ is minimal when the second term $(\mathbb{E}X-b)^2$ is minimal, i.e. when it is zero. This is obtained for the choice $b = \mathbb{E}X$.

Variance Properties

Theorem

Let X be a random variable with finite variance and a and b be real constants. Then,

1. $\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2$, *i.e.* $\mu_2 = m_2 - m_1^2$
2. $\mathbb{V}ar(aX + b) = a^2\mathbb{V}ar(X)$.

HW: Assume $X \sim Bin(n, p)$. Show that

$$\mathbb{V}ar(X) = np(1 - p)$$

by computing $\mathbb{E}(X) = np$ and $\mathbb{E}(X^2) = n(n - 1)p^2 + np$.

Skewness and kurtosis

Two quantities of interest, in addition to the mean and variance, are the skewness

$$\alpha_3 = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} \ ,$$

which measures the lack of symmetry in the pdf, and the kurtosis

$$\alpha_4 = \frac{\mu_4}{\mu_2^2} \ ,$$

which measures the peakedness or flatness of the pdf. (μ_k stands for the kth central moment of a random variable X)

1. Show that if a pdf is symmetric about a point a then $\alpha_3 = 0$.
2. Calculate α_3 for X with a pdf that is skewed to the right.

$$f_X(x) = \begin{cases} e^{-x}, & x \geq 0 \\ 0, & x < 0 \end{cases},$$

3. Calculate α_4 for each of the following pdfs and comment on the peakedness of each.

(1) $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}, \quad x \in \mathbb{R}$ (2) $f(x) = \frac{1}{2}, \quad -1 < x < 1.$

Moment Generating Function

The *moment generating function* (mgf) of a random variable X is a function associated with a probability distribution.
As its name suggests, the mgf can be used to generate moments of X.

Definition

Let X be a random variable with cdf F_X. The *moment generating function* (mgf) of X is

$$M_X(t) = \mathbb{E}(e^{tx}) = \begin{cases} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx & \text{if } X \text{ is cts} \\ \sum_x e^{tx} \mathbb{P}(X = x) & \text{if } X \text{ is discrete} \end{cases} \tag{6}$$

provided that the expectation exists for t in some neighborhood of 0.

If the expectation does not exist in a neighborhood of 0, then the mgf does not exist.

Poisson mgf

The mgf of a random variable X with a Poisson distribution $X \sim \mathcal{P}(\lambda)$ equals

$$\begin{aligned} M_X(t) = \mathbb{E}(e^{tX}) &= e^{-\lambda} \sum_{k=0}^{\infty} e^{tk} \cdot \frac{\lambda^k}{k!} \\ &= e^{-\lambda} \sum_{k=0}^{\infty} \frac{(\lambda\, e^t)^k}{k!} = e^{-\lambda} \cdot e^{\lambda\, e^t} \\ &= e^{\lambda\,(e^t - 1)} \end{aligned}$$

Gamma mgf

Let X be a random variable which follows a Gamma distribution with pdf

$$f(x) = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}}\, x^{\alpha-1}\, e^{-\frac{x}{\beta}}, \quad x > 0,\ \alpha, \beta > 0,$$

where $\Gamma(\alpha) = \int_0^{+\infty} x^{\alpha-1}\, e^{-x}\, dx$. Then the mgf of X is

$$\begin{aligned}
M_X(t) = \mathbb{E}(e^{tX}) &= \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \int_0^{+\infty} e^{tx} \cdot x^{\alpha-1}\, e^{-\frac{x}{\beta}}\, dx \\
&= \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \int_0^{+\infty} x^{\alpha-1} \cdot e^{-(\frac{1}{\beta}-t)\,x}\, dx \quad [\text{substitution: } (\frac{1}{\beta} - t)\,x = u] \\
&= \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \cdot \frac{1}{(\frac{1}{\beta}-t)^{\alpha}} \cdot \int_0^{+\infty} u^{\alpha-1}\, e^{-u} du = \frac{1}{\Gamma(\alpha)\,\beta^{\alpha}} \cdot \frac{1}{(\frac{1}{\beta}-t)^{\alpha}} \cdot \Gamma(\alpha) \\
&= \frac{1}{(1-\beta t)^{\alpha}},
\end{aligned}$$

when $t < \frac{1}{\beta}$. If $t \geq \frac{1}{\beta}$, the integral is not finite. Therefore, the mgf of a Gamma distribution exists if $t < \frac{1}{\beta}$.

Uniform mgf

Let $X \sim$ Uniform (a, b). Then,

$$M_X(t) = \mathbb{E}(e^{tX}) = \int_a^b \frac{e^{tx}}{b-a}\, dx = \frac{e^{bt} - e^{at}}{t(b-a)}.$$

Note that $\lim_{t \to 0} M_X(t) = 1$.

- [**HW**] Show that the mgf of a Binomial random variable $X \sim B(n, p)$ is of the form

$$M_X(t) = \left(pe^t + (1-p)\right)^n.$$

- [**HW**] Show that the mgf of an exponentially distributed random variable $X \sim \exp(\lambda)$, $\lambda > 0$, i.e.

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

  is of the form

$$M_X(t) = \frac{\lambda}{\lambda - t}, \qquad \text{for } t < \lambda.$$

Moments from mgf

Theorem

If X has mgf $M_X(t)$, then

$$E(X^k) = M_X^{(k)}(0) = \frac{d^k}{dt^k} M_X(t)\Big|_{t=0},$$

i.e. the kth moment of X is equal to the kth derivative of the mgf evaluated at $t = 0$.

Proof.

Assuming we can differentiate under the integral sign,

$$\begin{aligned}
\frac{d}{dt} M_X(t) &= \frac{d}{dt} \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\
&= \int_{-\infty}^{\infty} \frac{d}{dt} e^{tx} f_X(x) dx \\
&= \int_{-\infty}^{\infty} (x e^{tx}) f_X(x) dx \\
&= \mathbb{E}(X e^{tX})
\end{aligned}$$

so that $\frac{d^k}{dt^k} M_X(t)\Big|_{t=0} = \mathbb{E}(X e^{tX})\Big|_{t=0} = \mathbb{E}(X)$. Proceeding in an analogous manner for $k > 1$, we obtain the result. □

Moments from mgf

1. Consider $X \sim \mathcal{P}(\lambda)$. We have seen that $\mathbb{E}X = \mathbb{V}ar(X) = \lambda$.
   We can obtain these results by using the theorem.

$$\mathbb{E}(X) = M_X'(0) = \frac{d}{dt}\left(e^{\lambda\,(e^t-1)}\right)\Big|_{t=0} = e^{\lambda\,(e^t-1)} \cdot \lambda e^t\Big|_{t=0} = \lambda$$

$$\mathbb{E}(X^2) = M_X''(0) = \frac{d^2}{dt^2}\left(e^{\lambda\,(e^t-1)}\right)\Big|_{t=0} = e^{\lambda\,(e^t-1)}\left((\lambda e^t)^2 + \lambda e^t\right)\Big|_{t=0} = \lambda^2 + \lambda$$

   so that

$$\mathbb{V}ar(X) = \mathbb{E}(X^2) - (\mathbb{E}X)^2 = \lambda.$$

2. The first two moments of a Gamma distribution:

$$\mathbb{E}(X) = M_X'(0) = \frac{d}{dt}\left(\frac{1}{(1-\beta t)^\alpha}\right)\Big|_{t=0} = \frac{\alpha\beta}{(1-\beta t)^{\alpha+1}}\Big|_{t=0} = \alpha\beta$$

$$\mathbb{E}(X^2) = M_X''(0) = \frac{d^2}{dt^2}\left(\frac{1}{(1-\beta t)^\alpha}\right)\Big|_{t=0} = \frac{\alpha(\alpha+1)\beta^2}{(1-\beta t)^{\alpha+2}}\Big|_{t=0} = \alpha(\alpha+1)\beta^2$$

   and thus

$$\mathbb{V}ar(X) = E(X^2) - (EX)^2 = \alpha(\alpha+1)\beta^2 - \alpha^2\beta^2 = \alpha^2\beta^2.$$

Theorem

For any real constants a and b the mgf of the random variable $aX + b$ is given by

$$M_{aX+b}(t) = e^{bt}\, M_X(at).$$

Proof.

From the definition of mgf and properties of the exponential function we obtain

$$M_{aX+b}(t) = E\big(e^{(aX+b)t}\big) = E\big(e^{aXt} \cdot e^{bt}\big) = e^{bt} \cdot E(e^{aXt}) = e^{bt} \cdot M_{aX}(t),$$

provided that t is in a neighborhood of zero. □

mgf's *can* characterize distributions

Theorem

Let $F_X(x)$ and $F_Y(y)$ be two cdfs all of whose moments exist.

1. *If X and Y have bounded support then $F_X(u) = F_Y(u)$ for all u if and only if $E(X^r) = E(Y^r)$ for all $r \in \mathbb{N}_0$.*
2. *If the moment generating functions exist and $M_X(t) = M_Y(t)$ for all t in some neighborhood of 0, then $F_X(u) = F_Y(u)$ for all u.*

- If the mgf exists, it characterizes an infinite set of moments
- But, characterizing the set of moments is not enough to determine a distribution uniquely because there may be two distinct random variables with the same moments
- If a r.v. has bounded support or the mgf exists in a neighborhood of zero then the distribution is uniquely determined by the infinite set of its moments

Convergence of mgfs

Convergence of mgfs to a mgf implies convergence of cdfs.

Theorem

Let $(X_n)_{n\in\mathbb{N}}$ be a sequence of random variables, each with mgf $M_{X_n}(t)$, and let

$$\lim_{n\to\infty} M_{X_n}(t) = M_X(t), \quad \text{for all } t \text{ in a neighborhood of zero},$$

where $M_X(t)$ is a mgf. Then there exists a unique cdf F_X whose moments are determined by $M_X(t)$ and for all x where $F_X(x)$ is continuous,

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x).$$

Binomial-Poisson relationship

If there are many independent and identical Bernoulli experiments with a low probability of success, the number of successes can be approximated with a Poisson distribution. Namely, for $n \geq 10$, $p \leq \frac{1}{10}$ and $np \leq 10$ (Rule of thumb) a random variable $X \sim B(n, p)$ can be approximated by $Y \sim \mathcal{P}(\lambda)$, where $\lambda = np$, i.e.

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{(np)^k e^{-np}}{k!} = P(Y = k).$$

This can be proven by showing the convergence of the mgfs

$$M_X(t) = (pe^t + (1-p))^n \quad \to \quad e^{\lambda(e^r - 1)} = M_Y(t), \quad \text{as } n \to \infty.$$

Interchanging summation and differentiation

Theorem

Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ *converges for all* θ *in an interval* (a, b) *of real numbers and*

1. $\frac{\partial}{\partial \theta} h(\theta, x)$ *is continuous in* θ *for each* x,
2. $\sum_{x=0}^{\infty} h(\theta, x)$ *converges uniformly on every closed bounded subinterval of* (a, b).

Then,

$$\frac{\partial}{\partial \theta} \sum_{x=0}^{\infty} h(\theta, x) = \sum_{x=0}^{\infty} \frac{\partial}{\partial \theta} h(\theta, x)$$

The key condition to check is uniform convergence!

A series converges uniformly if its sequence of partial sums converges uniformly

Interchanging integration and summation

Theorem

Suppose that the series $\sum_{x=0}^{\infty} h(\theta, x)$ *converges for all* θ *in an interval* $[a, b]$ *of real numbers and that* $\frac{\partial}{\partial \theta} h(\theta, x)$ *is continuous in* θ *for each* x*. Then,*

$$\int_a^b \sum_{x=0}^{\infty} h(\theta, x) d\theta = \sum_{x=0}^{\infty} \int_a^b h(\theta, x) d\theta$$

Stein's lemma

Stein's lemma provides a way to calculate higher moments of a normal distribution.

Theorem

(Stein's lemma) Let $X \sim \mathcal{N}(\mu, \sigma^2)$ and let g be a differentiable function satisfying $E|g'(X)| < \infty$. Then,

$$\mathbb{E}\big(g(X)\,(X-\mu)\big) = \sigma^2\,\mathbb{E}g'(X) \tag{7}$$

(Higher order moments) Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then, by taking $g(x) = x^2$ we can calculate its third moment

$$\begin{aligned}\mathbb{E}(X^3) = \mathbb{E}X^2(X-\mu+\mu) &= \mathbb{E}\big(X^2\,(X-\mu)\big) + \mu\mathbb{E}(X^2) \\ &= \sigma^2\mathbb{E}(2X) + \mu(\mathbb{V}ar(X) + (\mathbb{E}X)^2) \\ &= 2\mu\sigma^2 + \mu(\sigma^2+\mu^2) = 3\mu\sigma^2 + \mu^3.\end{aligned}$$

Hwang's lemma

Similarly, Hwang's lemma gives a way to compute higher moments of a Poisson distribution.

Theorem

(Hwang's lemma) Let $g(x)$ be a function with finite $\mathbb{E}g(X)$ such that $g(-1)$ is also finite. If $X \sim \mathcal{P}(\lambda)$ then

$$\mathbb{E}\big(\lambda\, g(X)\big) = \mathbb{E}\big(X\, g(X-1)\big). \tag{8}$$

(Higher order moments) Let $X \sim \mathcal{P}(\lambda)$. Then, by taking $g(x) = x^2$ we can calculate its third moment. First, from (8) we obtain

$$\mathbb{E}\big(\lambda\, X^2\big) = \mathbb{E}(X\,(X-1)^2) = \mathbb{E}(X^3 - 2X^2 + X).$$

Then, the third moment of a Poisson distribution is

$$\mathbb{E}(X^3) = \lambda\,\mathbb{E}(X^2) + 2\mathbb{E}(X^2) - \mathbb{E}X = (\lambda+2)(\lambda^2+\lambda) - \lambda = \lambda^3 + 3\lambda^2 + \lambda.$$

[**HW**] Show the following analog to Stein's lemma, assuming appropriate conditions on the function g. If $X \sim Gamma(\alpha, \beta)$ then

$$\mathbb{E}\left(g(X)\,(X - \alpha\beta)\right) = \beta\mathbb{E}(X\, g'(X)) \tag{9}$$

Using (9), compute the third moment of X.

Characteristic Functions

The characteristic function of X is defined by

$$\phi_X(x)(t) = \mathbb{E}(e^{itX})$$

where i is the complex number $\sqrt{-1}$.

- When the momets of F_X exist, ϕ_X can be used to generate them, much like the mgf.
- The characteristic function always exists and it completely determines the distribution of X: every cdf has a unique characteristic function.

Theorem

(Convergence of Characteristic Functions) *Suppose $X_k, k = 1, 2, \ldots$ is a sequence of random variables, each with characteristic function $\phi_{X_k}(t)$. Furthermore, suppose*

$$\lim_{k\to\infty} \phi_{X_k}(t) = \phi_X(t), \quad \text{for all } t \text{ in a neighborhood of } 0,$$

and $\phi_X(t)$ is a characteristic function. Then, for all x where $F_X(x)$ is continuous,

$$\lim_{k\to\infty} F_{X_k}(x) = F_X(x)$$