Transformations and Expectations

Ricardo Dahis

PUC-Rio, Department of Economics

Summer 2023

Contents

Contents

functions of a random variable

- if X is a random variable with cdf $F_X(x)$, than any function $Y = g(X)$ is also a random variable
- we write $y = g(x)$, with

$$g(x) \quad : \quad \Omega_X \to \Omega_Y$$

- since Y is a function of X, we can describe the probabilistic behavior of Y in terms of X

$$\mathbb{P}(Y \in A) \quad = \quad \mathbb{P}(g(X) \in A) \;\; \text{for any event} \;\; A$$

- we associate g with an inverse mapping, denoted g^{-1},

$$\begin{aligned} g^{-1}(A) &= \{x \in \Omega_X : g(x) \in A\} \\ g^{-1}(\{y\}) &= \{x \in \Omega_X : g(x) = y\} \end{aligned}$$

  where g^{-1} is usually a **set**.

- **remark**: we only write $g^{-1}(y) = x$ if $\exists! x$ for which $g(x) = y$

probability distribution of Y

- if $Y = g(X)$, then we can write for any set $A \subset \Omega_Y$

$$\begin{aligned} \mathbb{P}(Y \in A) &= \mathbb{P}(g(X) \in A) \\ &= \mathbb{P}(x \in \Omega_X : g(x) \in A) \\ &= \mathbb{P}(X \in g^{-1}(A)) \end{aligned}$$

  which satisfies Kolmogorov's axioms

- if X is discrete, then the sample space Ω_X is countable.
  - the sample space for $Y = g(X)$ is $\Omega_Y = \{y : y = g(x), x \in \Omega_X\}$ is also countable
  - it follows that Y is also a discrete random variable
  - The pmf for Y is $\forall y \in S_Y$,

$$f_Y(y) = \mathbb{P}(Y = y) = \sum_{x \in g^{-1}(y)} \mathbb{P}(X = x) = \sum_{x \in g^{-1}(y)} f_X(x)$$

binomial rv

- **definition**: a discrete random variable X has a binomial distribution if its pmf is of the form

$$f_X(x) \;=\; \mathbb{P}(X = x) \;=\; \binom{n}{x} p^x (1-p)^{n-x}$$

for $x = 0, 1, \ldots, n$ (with $0 < p < 1$) $X \sim \mathrm{Bin}(n, p)$

- example: what is the probability of obtaining exactly three heads in five coin tosses? Suppose that the probability of a head is $\frac{1}{4}$.

$$P(X = 3) \;=\; \binom{5}{3} \left(\frac{1}{4}\right)^3 \left(1 - \frac{1}{4}\right)^2 \;\approx\; 8.79\%$$

where the binomial coefficient accounts for rearrengements of the sequence of heads and tails.

binomial transformation

what is the probability of getting exactly two tails in the example above?

- consider $Y = n - X$. $\Omega_Y = \{y : y = g(x),\, x \in \Omega_X\} = \{0, 1, \ldots, n\}$
- since $g^{-1}(y)$ is a single point $\{x = n - y\}$,

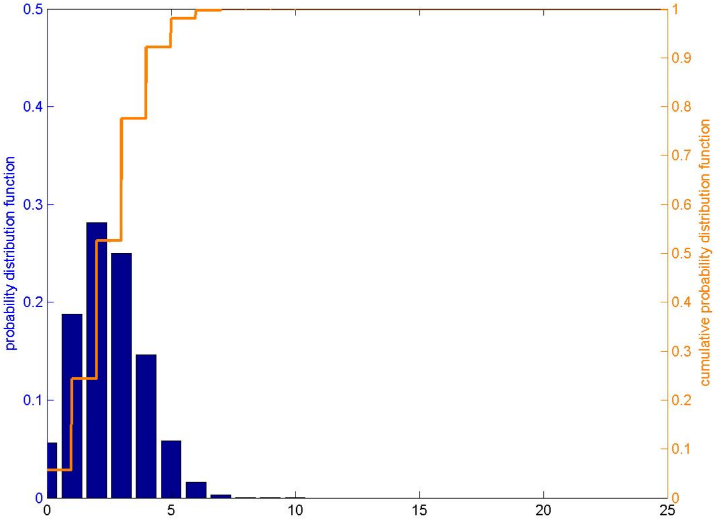
$$\begin{aligned} f_Y(y) &= \sum_{x \in g^{-1}(y)} f_X(x) = f_X(n-y) \\ &= \binom{n}{n-y} p^{n-y}(1-p)^y = \binom{n}{y}(1-p)^y p^{n-y} \end{aligned}$$

$Y \sim \mathrm{Bin}(n, 1-p)$

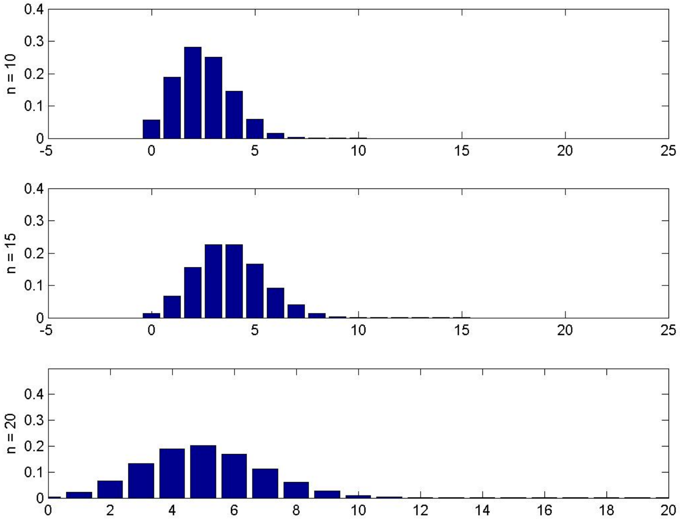
- the last equality follows from the fact that

$$\binom{n}{n-y} = \frac{n!}{(n-n+y)!(n-y)!} = \frac{n!}{y!(n-y)!} = \binom{n}{y}$$

looks of the binomial distribution



binomial distribution as n grows



functions of a continuous random variable

- it is sometimes possible to find simple formulae for the cdf and pdf of $Y = g(X)$ in terms of the cdf and pdf of X and the function g

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(Y \leq y) \\
&= \mathbb{P}(g(X) \leq y) \\
&= \mathbb{P}\left(\{x \in \Omega_X : g(x) \leq y\}\right) \\
&= \int_{\{x \in \Omega_X : g(x) \leq y\}} f_X(x)\,\mathrm{d}x
\end{aligned}
$$

- **though...** identifying $\{x \in \Omega_X : g(x) \leq y\}$ and integrating $f_X(x)$ over this region are not always easy

uniform transformation

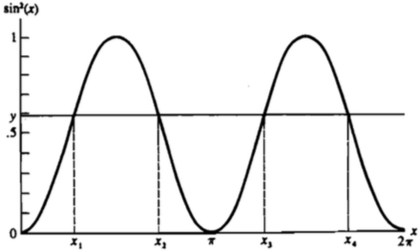
- **definition**: a continuous random variable X with pdf of the form

$$f_X(x) = \frac{1}{b-a}$$

  for $a < x < b$ (and zero otherwise) is said to have a uniform distribution between (a, b)
  $X \sim \mathrm{U}(a, b)$

- consider $X \sim \mathrm{U}(0, 2\pi)$ and $Y = \sin^2(X)$, then
  - $S_Y = \{y : y = \sin^2(x), x \in (0, 2\pi)\} = [0, 1]$
  - $g^{-1}(y)$ is not a single point.

uniform transformation



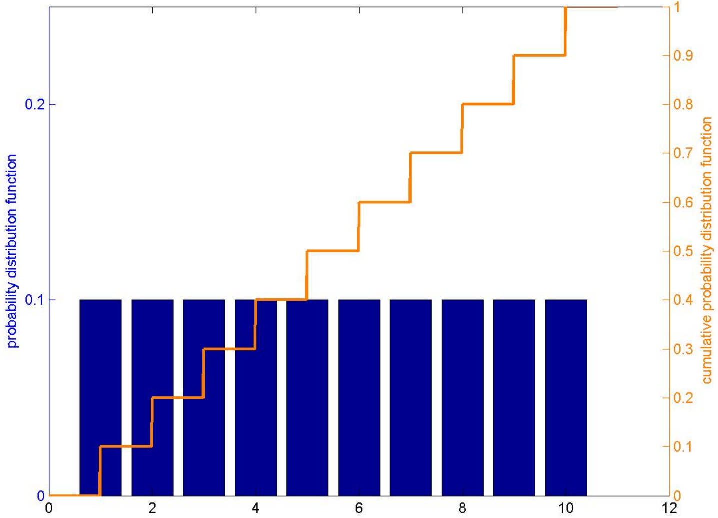
- Then

$$\begin{aligned}\mathbb{P}(Y \leq y) &= \mathbb{P}(X \leq x_1) + \mathbb{P}(x_2 \leq X \leq x_3) + \mathbb{P}(X \geq x_4) \\ &= 2\mathbb{P}(X \leq x_1) + 2\mathbb{P}(x_2 \leq X \leq \pi),\end{aligned}$$

  due to symmetry of $\sin^2(\cdot)$ and uniformity of X

- Even in this apparently simple case, the expression for the cdf of Y is not that simple...

looks of the uniform distribution



keeping track of the sample space

- it is important to keep track of the support of the distribution:

$$
\begin{aligned}
\Omega_X &= \{x : f_X(x) > 0\} \\
\Omega_Y &= \{y : y = g(x) \text{ for some } x \in S_X\}
\end{aligned}
$$

- it is easiest to deal with monotone transformations, because then it is a one-to-one mapping from Ω_X onto Ω_Y, uniquely pairing (x, y)
  - increasing if $u > v \Rightarrow g(u) > g(v)$
  - decreasing if $u > v \Rightarrow g(u) < g(v)$

- if g is monotone, $g(\cdot)$ is a bijection and $g^{-1}(y)$ is single-valued, that is, $g^{-1}(y) = x$ if and only if $y = g(x)$

keeping track of the sample space

- if increasing,

$$\begin{aligned}\{x \in \Omega_X : g(x) \leq y\} &= \{x \in \Omega_X : g^{-1}(g(x)) \leq g^{-1}(y)\} \\ &= \{x \in \Omega_X : x \leq g^{-1}(y)\}\end{aligned}$$

  then

$$F_Y(y) = \int_{x \in \Omega_X : x \leq g^{-1}(x)} f_X(x)dx = \int_{-\infty}^{g^{-1}(x)} f_X(x)dx = F_X(g^{-1}(y))$$

keeping track of the sample space

- if decreasing,

$$\begin{aligned}\{x \in \Omega_X : g(x) \leq y\} &= \{x \in \Omega_X : g^{-1}(g(x)) \geq g^{-1}(y)\} \\ &= \{x \in \Omega_X : x \geq g^{-1}(y)\}\end{aligned}$$

  then

$$F_Y(y) = \int_{x \in \Omega_X : x \geq g^{-1}(x)} f_X(x)dx = \int_{g^{-1}(x)}^{\infty} f_X(x)dx = 1 - F_X(g^{-1}(y))$$

cdf of a monotone transformation

example: uniform-exponential relationship.

- Suppose that $X \sim \mathrm{U}(0,1)$, for which $f_X(x) = 1$ and $F_X(x) = x$, and then make the transformation $Y = g(X) = -\ln X$

- Since

$$\frac{d}{dx} g(x) \quad = \quad \frac{d}{dx}(-\ln X) \quad = \quad -\frac{1}{x} < 0 \quad \text{for} \quad 0 < x < 1$$

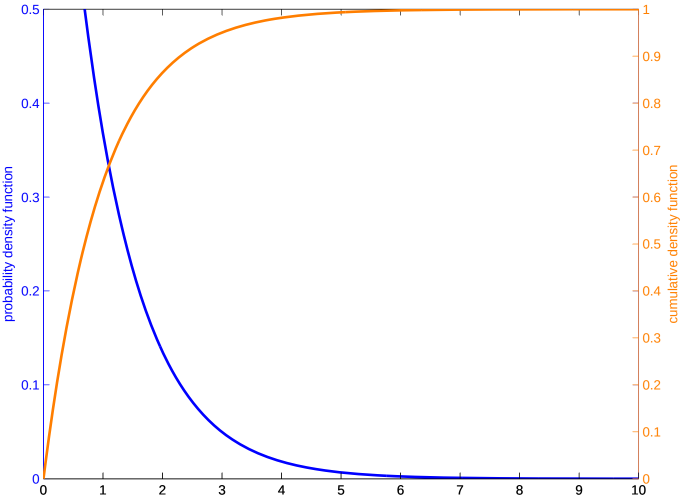
  and the transformation is monotone and decreasing.

- $S_X = (0,1) \quad \Rightarrow \quad S_Y = (0,\infty)$

- for any $y > 0$, $x = e^{-y}$ and hence $F_Y(y) = 1 - F_X(e^{-y}) = 1 - e^{-y}$

(exponential distribution)

looks of the exponential distribution



pdf of a monotone transformation

- if $Y = g(X)$ is a continuous random variable, then we may obtain its pdf by differentiating the cdf

- **theorem**: let $X \sim f_X(x)$ with support Ω_X and $Y = g(X)$, where g is monotone; if $f_X(x)$ is continuous on Ω_X and $g^{-1}(y)$ has a continuous derivative on Ω_Y, then the pdf of Y is given by

$$f_Y(y) \quad = \quad \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| & y \in S_Y \\ 0 & \text{otherwise} \end{cases}$$

- **proof**: we have that, by the chain rule,

$$f_Y(y) \quad = \quad \frac{d}{dy} F_Y(y) \quad = \quad \begin{cases} f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) & \text{if increasing} \\ \\ -f_X(g^{-1}(y)) \frac{d}{dy} g^{-1}(y) & \text{if decreasing} \end{cases}$$

which can be expressed as above. ■

example

- term $\frac{d}{dy}g^{-1}(y)$ is an indirect adjustment for the support of the distribution
- take the simplest example: $X \sim U([0,1])$, therefore $f_X(x) = 1$
- the transformation $g(x) = 2x$ produces $Y \sim U([0,2])$ with $f_Y(y) = \frac{1}{2}$
- using the proof above,

$$f_Y(y) \quad = \quad \underbrace{f_X(g^{-1}(y))}_{=1} \cdot \underbrace{\frac{d}{dy}g^{-1}(y)}_{=\frac{1}{2}} \quad = \quad \frac{1}{2}$$

i.e., the distribution "thinned out" over a larger support set

inverted gamma

- **example:** let $X \sim \mathrm{G}(n, \beta)$, with pdf

$$f_X(x) = \frac{1}{(n-1)!\beta^n} x^{n-1} e^{-x/\beta}, \qquad 0 < x < \infty$$

  where β is a positive constant, n is a positive integer, and $Y = 1/X$

- then $\Omega_X = \Omega_Y = (0, \infty)$ and $g^{-1}(y) = 1/y$, yielding

$$\begin{aligned} f_Y(y) &= f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| \\ &= \frac{1}{(n-1)!\beta^n} y^{1-n} e^{-1/(\beta y)} \frac{1}{y^2} \\ &= \frac{1}{(n-1)!\beta^n} y^{-(n+1)} e^{-1/(\beta y)} \qquad \text{(inverted gamma)} \end{aligned}$$

piecewise monotone transformations

- it is sometimes the case that g is monotone over **certain intervals**, allowing us to apply the above results for each one of these regions

- **example**: square transformation $Y = X^2$ for a continuous X
  - $F_Y(y) = \mathbb{P}(X^2 \leq y) = \mathbb{P}(-\sqrt{y} \leq X \leq \sqrt{y}) = F_X(\sqrt{y}) - F_X(-\sqrt{y})$
  - $f_Y(y) = \frac{d}{dy} F_Y(y) = \frac{1}{2\sqrt{y}} \left[f_X(\sqrt{y}) + f_X(-\sqrt{y}) \right]$ for $y > 0$
  - expression is the sum of two pieces that represent the intervals where $g(x) = x^2$ is monotone.

more formally...

- **theorem** (CB 2.1.8): let X have pdf $f_X(x)$, let $Y = g(X)$, and suppose that there exists a partition $A_0, A_1, \ldots, A_k$ of Ω_X such that $\mathbb{P}(X \in A_0) = 0$ and $f_X(x)$ is continuous on each A_i, and that there exist functions $g_1(x), \ldots, g_k(x)$ defined respectively on $A_1, \ldots, A_k$ that satisfy

  (i) $g(x) = g_i(x)$ for $x \in A_i$

  (ii) $g_i(x)$ is monotone on A_i

  (iii) the set $\Omega_Y = \{y : y = g_i(x), x \in A_i\}$ is the same for each $i = 1, \ldots, k$

  (iv) g_i^{-1} has a continuous derivative on S_Y for each $i = 1, \ldots, k$

  then...

$$f_Y(y) = \begin{cases} \sum_{i=1}^{k} f_X(g_i^{-1}(y)) \left| \frac{d}{dy} g_i^{-1}(y) \right| & y \in \Omega_Y \\ 0 & \text{otherwise} \end{cases}$$

chi-squared distribution

- **example**: let $X \sim N(0, 1)$, with pdf

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

and consider $g(x) = x^2$, which is monotone on $(-\infty, 0)$ and $(0, \infty)$.

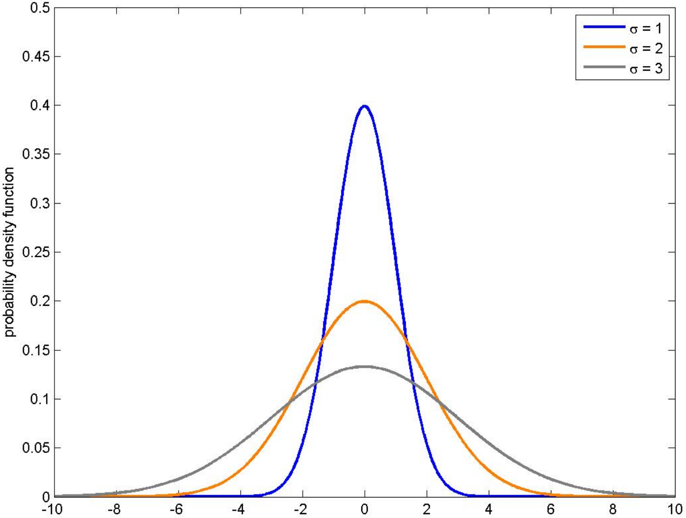
- According to the theorem above, we choose

$$\begin{aligned} A_0 &= \{0\} \\ A_1 &= (-\infty, 0), \quad g_1(x) = x^2, \quad g_1^{-1}(y) = -\sqrt{y} \\ A_2 &= (0, \infty), \quad g_2(x) = x^2, \quad g_2^{-1}(y) = \sqrt{y} \end{aligned}$$

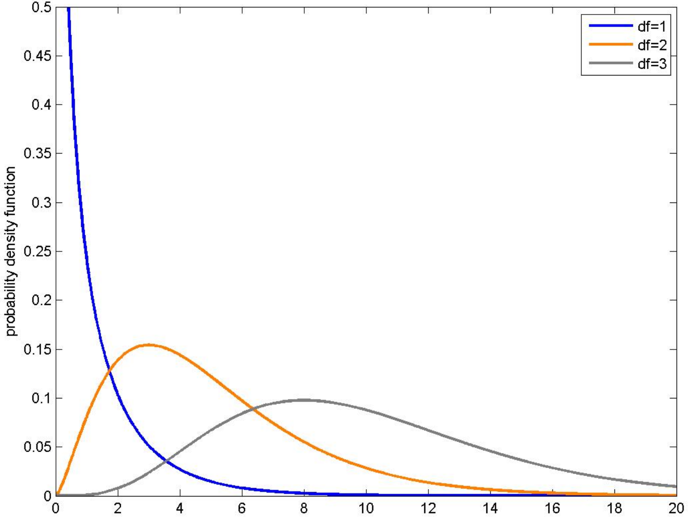
$$\begin{aligned} f_Y(y) &= \frac{1}{2\sqrt{y}} \left[\phi(\sqrt{y}) + \phi(-\sqrt{y})\right] \\ &= \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{y}} e^{-y/2}, \qquad 0 < y < \infty \end{aligned}$$

(chi-squared distribution, 1 dof)

looks of the normal distribution



looks of the chi-squared distribution



probability integral transform

- **theorem**: let X have a continuous and strictly increasing cdf $F_X(x)$ and define the random variable $Y = F_X(X)$. Then $Y \sim \mathrm{U}(0,1)$
- **proof**: for $0 < y = F_X(x) < 1$,

$$
\begin{aligned}
F_Y(y) &= \mathbb{P}(F_X(X) \le y) \\
&= \mathbb{P}\left(F_X^{-1}[F_X(X)] \le F_X^{-1}(y)\right) \\
&= \mathbb{P}\left(X \le F_X^{-1}(y)\right) \\
&= F_X\left(F_X^{-1}(y)\right) = y \quad \Leftrightarrow \quad y \sim \mathrm{U}(0,1)
\end{aligned}
$$

$\blacksquare$

- **remark 1**: defining $F_X^{-1}(y) = \inf\{x : F_X(x) \ge y\}$ for $0 < y < 1$ avoids constancy issues in F_X
- **remark 2**: very useful theorem in it allows generating random numbers from all distributions. Sample u from $U([0,1])$ and solve $F_X(x) = u$.

Contents

expected value, mean, expectation,...

- **definition**: the expected value of a random variable $g(X)$ is

$$\mathbb{E}g(X) = \begin{cases} \int_{-\infty}^{\infty} g(x)\, f_X(x)\, \mathrm{d}x & \text{if continuous} \\ \sum_{x \in S_X} g(x) f_X(x) = \sum_{x \in S_X} g(x) \mathbb{P}(X = x) & \text{if discrete} \end{cases}$$

  if infinite, we say that the expectation does not exist.

- **example**: exponential mean
  - suppose $X \sim \text{Exp}(\lambda)$, with pdf $f_X(x) = \frac{1}{\lambda}\, e^{-x/\lambda}$ for $\lambda > 0$ and $x \geq 0$

$$\begin{aligned} \mathbb{E}(X) &= \int_0^{\infty} \frac{x}{\lambda}\, e^{-x/\lambda}\, \mathrm{d}x \\ &= \left(-xe^{-x/\lambda}\right)\Big|_0^{\infty} + \int_0^{\infty} e^{-x/\lambda}\, \mathrm{d}x \\ &= \int_0^{\infty} e^{-x/\lambda}\, \mathrm{d}x \;=\; \lambda \end{aligned}$$

    since $\int_0^{\infty} \frac{1}{\lambda} e^{-x/\lambda} = 1$.

mean of a binomial random variable

- if $X \sim \text{Bin}(n, p)$, then

$$
\begin{aligned}
\mathbb{E}(X) &= \sum_{x=0}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=1}^{n} x \binom{n}{x} p^x (1-p)^{n-x} \\
&\overset{\star}{=} n \sum_{y=0}^{n-1} \binom{n-1}{y} p^{y+1} (1-p)^{n-y-1} \\
&= np \underbrace{\sum_{y=0}^{n-1} \binom{n-1}{y} p^{y} (1-p)^{(n-1)-y}}_{=1} \\
&= np
\end{aligned}
$$

equality ($\star$) used the fact that $x\binom{n}{x} = n\binom{n-1}{x-1}$.

mean of the Cauchy distribution

- if $X \sim$ Cauchy with pdf $f_X(x) = \frac{1}{\pi(1+x^2)}$, then

$$
\begin{aligned}
\mathbb{E}|X| &= \int_{-\infty}^{\infty} \frac{|x|}{\pi} \frac{1}{1+x^2}\,\mathrm{d}x \\
&= \frac{2}{\pi} \int_0^{\infty} \frac{x}{1+x^2}\,\mathrm{d}x \\
&= \lim_{M\to\infty} \frac{2}{\pi} \int_0^{M} \frac{x}{1+x^2}\,\mathrm{d}x \\
&= \lim_{M\to\infty} \frac{2}{\pi} \left.\frac{\ln(1+x^2)}{2}\right|_0^M \\
&= \lim_{M\to\infty} \frac{2}{\pi} \frac{\ln(1+M^2)}{2} \quad = \quad \infty
\end{aligned}
$$

however... if $\mathbb{E}|X| = \infty$, then $\mathbb{E}(X) = \infty$ as well!!

expectation as a linear operator

- **taking expectations is a linear operation** because, for any random variables X and Y,

$$\mathbb{E}(aX + bY) = a\mathbb{E}(X) + b\mathbb{E}(Y)$$

  as long as a and b are constants

- moreover (CB 2.2.5), if the expectations of $g(x)$ and $h(x)$ exists,
  - (a) $\mathbb{E}[ag(X) + bh(X) + c] = a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)] + c$
  - (b) if $g(x) \geq 0$ for all x, then $\mathbb{E}[g(X)] \geq 0$
  - (c) if $g(x) \geq h(x)$ for all x, then $\mathbb{E}[g(X)] \geq \mathbb{E}[h(X)]$
  - (d) if $a \leq g(x) \leq b$ for all x, then $a \leq \mathbb{E}[g(X)] \leq b$

proofs for a continuous random variable

(a) by definition,

$$\begin{aligned}
\mathbb{E}[ag(X) + bh(X) + c] &= \int_{-\infty}^{\infty} [ag(x) + bh(x) + c] f_X(x)\,\mathrm{d}x \\
&= a \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x + b \int_{-\infty}^{\infty} h(x) f_X(x)\,\mathrm{d}x \\
&\quad + c \int_{-\infty}^{\infty} f_X(x)\,\mathrm{d}x \\
&= a\mathbb{E}[g(X)] + b\mathbb{E}[h(X)] + c \qquad \blacksquare
\end{aligned}$$

(b, c, d) $\mathbb{E}[\ell(X)] = \int_{-\infty}^{\infty} \ell(x) f_X(x)\,\mathrm{d}x \geq 0$ given that $\ell(x) \geq 0$ and that $f_X(x) > 0$ for $x \in S_X$ $\blacksquare$

expectation as distance minimizer

- **theorem** (CB 2.2.6): measure the distance between a random variable X and a constant μ by the mean quadratic distance $\mathbb{E}(X-\mu)^2$. Then $\mu = \mathbb{E}(X)$ is the value that minimizes that distance

- **proof**: algebraic, rather than the usual calculus approach

$$\begin{aligned}
\mathbb{E}(X-\mu)^2 &= \mathbb{E}[X - \mathbb{E}(X) + \mathbb{E}(X) - \mu]^2 \\
&= \mathbb{E}[X - \mathbb{E}(X)]^2 + \mathbb{E}[\mathbb{E}(X) - \mu]^2 + 2\mathbb{E}[X - \mathbb{E}(X)][\mathbb{E}(X) - \mu] \\
&= \mathbb{E}[X - \mathbb{E}(X)]^2 + [\mathbb{E}(X) - \mu]^2 + 2[\mathbb{E}(X) - \mathbb{E}(X)][\mathbb{E}(X) - \mu] \\
&= \mathbb{E}[X - \mathbb{E}(X)]^2 + [\mathbb{E}(X) - \mu]^2
\end{aligned}$$

  both terms are nonnegative and, although we have no control over the variance of X, the second term is zero for $\mu = \mathbb{E}(X)$ $\blacksquare$

Contents

uncentered vs centered moments

- **definition**: the kth (uncentered) moment of X is

$$\tilde{\mu}_k = \mathbb{E}(X^k)$$

whereas the kth centered (or central) moment is

$$\mu_k = \mathbb{E}(X-\mu)^k$$

where $\mu = \tilde{\mu}_1 = \mathbb{E}(X)$

- **examples**:
  - dispersion $\sim$ variance $\qquad \sigma^2 = \mathrm{var}(X) = \mathbb{E}(X-\mu)^2$
  - asymmetry $\sim$ skewness $\qquad \mathrm{sk}(X) = \mathbb{E}[(X-\mu)/\sigma]^3$
  - tail thickness $\sim$ kurtosis $\qquad \mathrm{k}(X) = \mathbb{E}[(X-\mu)/\sigma]^4$

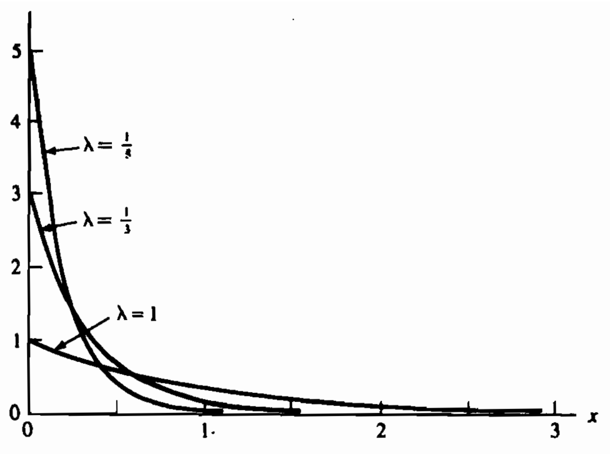
exponential variance

- **example**: let $X \sim \text{Exp}(\lambda)$ and hence $\mathbb{E}(X) = \lambda$

$$
\begin{aligned}
\text{var}(X) &= \mathbb{E}(X-\lambda)^2 \\
&= \int_0^\infty (x-\lambda)^2 \, \frac{1}{\lambda} \, e^{-x/\lambda} \, \mathrm{d}x \\
&= \int_0^\infty (x^2 - 2x\lambda + \lambda^2) \, \frac{1}{\lambda} \, e^{-x/\lambda} \, \mathrm{d}x \\
&= \int_0^\infty x^2 \, \frac{1}{\lambda} \, e^{-x/\lambda} \, \mathrm{d}x - 2\lambda \int_0^\infty x \, \frac{1}{\lambda} \, e^{-x/\lambda} \, \mathrm{d}x + \lambda^2 \int_0^\infty \frac{1}{\lambda} \, e^{-x/\lambda} \, \mathrm{d}x \\
&= 2\lambda^2 - 2\lambda^2 + \lambda^2 \\
&= \lambda^2
\end{aligned}
$$

- overdispersion coefficient $= \dfrac{\text{standard deviation}}{\text{mean}} = 1$

shape of the exponential distribution



variance of an affine transformation

- **theorem**: if X is a random variable with finite variance, it then follows that

$$\text{var}(aX + b) = a^2\text{var}(X)$$

  for any constants a and b

- **proof**: it follows from the definition that

$$\begin{aligned} \text{var}(aX + b) &= \mathbb{E}[aX + b - \mathbb{E}(aX + b)]^2 \\ &= \mathbb{E}[aX + b - a\mathbb{E}(X) - b]^2 \\ &= \mathbb{E}[aX - a\mathbb{E}(X)]^2 \\ &= a^2\,\mathbb{E}[X - \mathbb{E}(X)]^2 \\ &= a^2\,\text{var}(X) \end{aligned}$$

- it is sometimes easier to use $\text{var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2$

$$\begin{aligned} \text{var}(X) &= \mathbb{E}[X - \mathbb{E}X]^2 \\ &= \mathbb{E}[X^2 - 2X\mathbb{E}X + (\mathbb{E}X)^2] \\ &= \mathbb{E}X^2 - (\mathbb{E}X)^2 \end{aligned}$$

binomial variance

- **example**: let $X \sim \mathrm{Bin}(n, p)$ and hence $\mathbb{E}(X) = np$

$$\begin{aligned}
\mathbb{E}(X^2) &= \sum_{x=0}^{n} x^2 \binom{n}{x} p^x (1-p)^{n-x} = \sum_{x=0}^{n} xn \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= n \sum_{x=1}^{n} x \binom{n-1}{x-1} p^x (1-p)^{n-x} \\
&= n \sum_{y=0}^{n-1} (y+1) \binom{n-1}{y} p^{y+1} (1-p)^{(n-1)-y} \\
&= np \sum_{y=0}^{n-1} y \binom{n-1}{y} p^y (1-p)^{(n-1)-y} + np \sum_{y=0}^{n-1} \binom{n-1}{y} p^y (1-p)^{(n-1)-y} \\
&= np[(n-1)p + 1]
\end{aligned}$$

- $\mathrm{var}(X) = \mathbb{E}(X^2) - [\mathbb{E}(X)]^2 = np(np - p + 1) - (np)^2 = np(1-p)$

moment generating function

- **definition**: let X be a random variable with cdf F_X; the moment generating function (mgf) of X is given by

$$M_X(t) = \mathbb{E}(e^{tX})$$

  if there is an $h > 0$ such that $\mathbb{E}(e^{tX})$ exists for all t in $-h < t < h$. If the expectation does not exist in a neighborhood of 0, we say that the moment generating function does not exist.

- More explicitly,

$$\begin{aligned} M_X(t) &= \int_{-\infty}^{\infty} e^{tx} f_X(x) dx \\ M_X(t) &= \sum_x e^{tx} P(X = x) \end{aligned}$$

  if X is continuous and discrete, respectively.

moment generating function

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

$$\mathbb{E}(X^k) = M_X^{(k)}(0) \quad \text{with} \quad M_X^{(k)}(0) = \frac{\mathrm{d}^k}{\mathrm{d}t^k} M_X(t)\bigg|_{t=0}$$

- **proof**: assuming that we may differentiate under the integral

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} M_X(t) &= \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} e^{tx} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} \frac{\mathrm{d}}{\mathrm{d}t} e^{tx} f_X(x)\,\mathrm{d}x \\
&= \int_{-\infty}^{\infty} x e^{tx} f_X(x)\,\mathrm{d}x = \mathbb{E}(X e^{tX}) \\
M_X^{(1)}(0) &= \mathbb{E}\left(X e^{tX}\right)\Big|_{t=0} = \mathbb{E}(X)
\end{aligned}$$

and, analogously, $M_X^{(k)}(0) = \mathbb{E}\left(X^k e^{tX}\right)\big|_{t=0} = \mathbb{E}(X^k)$ ■

gamma mgf

- **example**: let $X \sim \mathrm{G}(\alpha, \beta)$ with pdf

$$
\begin{aligned}
f_X(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} e^{-x/\beta}, \qquad 0 < x < \infty \qquad \alpha > 0 \qquad \beta > 0 \\
M_X(t) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} \int_0^\infty x^{\alpha-1} e^{-x(1/\beta - t)}\, \mathrm{d}x \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \underbrace{\int_0^\infty x^{\alpha-1} e^{-x\left(\frac{\beta}{1-\beta t}\right)^{-1}}\, \mathrm{d}x}_{= \int_0^\infty x^{a-1} e^{-x/b} dx \; = \; \Gamma(a) b^a} \\
&= \frac{1}{\Gamma(\alpha)\beta^\alpha} \Gamma(\alpha) \left(\frac{\beta}{1-\beta t}\right)^\alpha \\
&= \left(\frac{1}{1-\beta t}\right)^\alpha, \qquad \text{if } t < 1/\beta \\
\mathbb{E}(X) &= M_X^{(1)}(0) = \left.\frac{\alpha\beta}{(1-\beta t)^{\alpha+1}}\right|_{t=0} = \alpha\beta
\end{aligned}
$$

- if $t \geq 1/\beta$ the mgf does not exist.

standard normal mgf

- **example:** let X have the standard normal distribution. Then

$$M_X(t) \quad = \quad \mathbb{E}(e^{tX}) \;=\; \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \;=\; \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}+tx} dx$$

and we have that

$$-\frac{x^2}{2} + tx \quad = \quad \frac{1}{2}t^2 - \frac{1}{2}(x-t)^2$$

and so

$$M_X(t) \quad = \quad e^{\frac{1}{2}t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-t)^2}{2}} dx \;=\; e^{\frac{1}{2}t^2}$$

binomial mgf

- **example**: let $X \sim \text{Bin}(n, p)$, so the moment generating function is

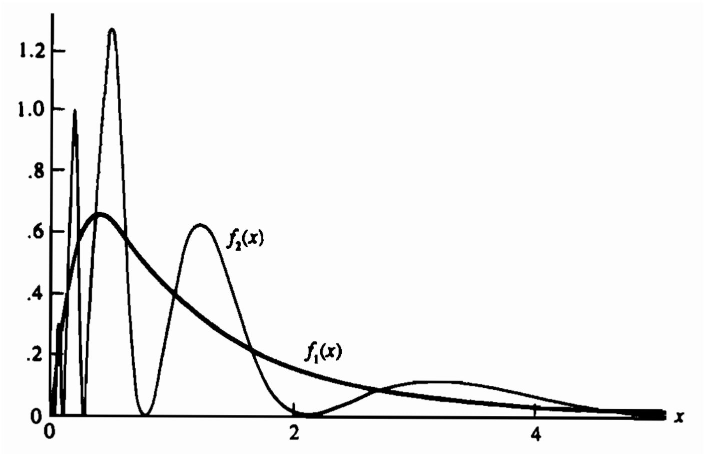
$$
\begin{aligned}
M_X(t) &= \sum_{x=0}^{n} e^{tx} \binom{n}{x} p^x (1-p)^{n-x} \\
&= \sum_{x=0}^{n} \binom{n}{x} (pe^t)^x (1-p)^{n-x} \\
&= \left[pe^t + (1-p)\right]^n
\end{aligned}
$$

  given that the binomial theorem yields $(u+v)^n = \sum_{x=0}^{n} \binom{n}{x} u^x v^{n-x}$ (proof left as exercise)

why mgf?

- usefulness of the mgf:
  - if mgf exists, yields infinite number of moments;
  - and especially so when moments characterize distributions.
- this begs the question: are two distributions equivalent if they have the same infinite set of moments?
- unfortunately, the answer is *no*: there may exist two different distributions with the same infinite moments.

two pdfs with the same moments



a counterexample

- let Z be a standard normal with p.d.f. $\phi(z)$ and $x = e^z \Leftrightarrow z = \ln x$ with $x \in (0, \infty)$ and $z \in \mathbb{R}$. Then

$$f(x) \quad = \quad \frac{\phi(\ln x)}{x} \quad = \quad \frac{1}{\sqrt{2\pi}x} e^{\frac{-(\ln x)^2}{2}}$$

and then $\mathbb{E}(X^k) = \mathbb{E}(e^{kZ}) = e^{\frac{k^2}{2}}$.

- now let $g(x) = f(x)[1 + h(x)]$, where $h(x) = \sin(2\pi \ln x)$. Then

$$\int_0^\infty x^k f(x) h(x) dx \quad = \quad \int_0^\infty x^k \frac{1}{\sqrt{2\pi}x} e^{\frac{-(\ln x)^2}{2}} \sin(2\pi \ln x) dx.$$

a counterexample

- let U have a normal distribution with mean 0 and unitary variance. Writing $u = \ln x$, we get

$$\begin{aligned}
\int_0^\infty e^{uk} \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}} \sin(2\pi u) du &= \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-u^2}{2}+uk} \sin(2\pi u) du \\
&= e^{\frac{1}{2}k^2} \int_0^\infty \frac{1}{\sqrt{2\pi}} e^{\frac{-(u-k)^2}{2}} \sin(2\pi u) du \\
&= e^{\frac{1}{2}k^2} \mathbb{E}\left[\sin(2\pi u)\right] = 0
\end{aligned}$$

  where the last equality follows from symmetry of $\sin(\cdot)$.

- it follows that

$$\begin{aligned}
\int_0^\infty x^k g(x) dx &= \int_0^\infty x^k f(x) dx + \int_0^\infty x^k f(x) h(x) dx \\
&= \int_0^\infty x^k f(x) dx
\end{aligned}$$

  so $f(x)$ and $g(x)$ have the same moments.

- in other words: two different distributions may have the same infinite set of moments ☹

how to circumvent non-uniqueness of moments?

- problem does not arise for random variables with a bounded support. In this case, infinite sequence of moments does uniquely determine distribution
- **theorem**: if F_X and F_Y have all moments, then...
  (a) if X and Y have bounded support, then $F_X(u) = F_Y(u)$ for all u if and only if $\mathbb{E}(X^k) = \mathbb{E}(Y^k)$ for all integers $k = 0, 1, 2, \ldots$

  (b) if $M_X(t) = M_Y(t)$ for all t in some neighborhood of zero, then $F_X(u) = F_Y(u)$ for all u.

log-normal counterexample

- example: log-normal mgf doesn't exist, so it can't fulfill conditions of previous theorem.

$$
\begin{aligned}
\mathbb{E}(e^{tX}) &= \mathbb{E}\left[\sum_{k=0}^{\infty} \frac{(tX)^k}{k!}\right] = \sum_{k=0}^{\infty} t^k \frac{\mathbb{E}(X^k)}{k!} \\
&= \sum_{k=0}^{\infty} t^k \frac{e^{k^2/2}}{k!} = \infty
\end{aligned}
$$

for $t > 0$. Another way to see this is

$$
\begin{aligned}
\mathbb{E}(e^{tX}) &= \mathbb{E}(e^{te^Z}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{te^z} e^{-\frac{z^2}{2}} dz = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{te^z - \frac{z^2}{2}} dz \\
&\geq \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{te^z - \frac{z^2}{2}} dz \geq \frac{1}{\sqrt{2\pi}} \int_{0}^{\infty} e^{t\left[1+z+\frac{z^2}{2!}+\frac{z^3}{3!}\right] - \frac{z^2}{2}} dz = \infty
\end{aligned}
$$

convergence of mgfs

- **theorem**: suppose $\{X_n,\ n = 1, 2, \ldots\}$ is a sequence of random variables, each with mgf $M_{X_n}(t)$, such that

$$\lim_{n\to\infty} M_{X_n}(t) \quad = \quad M_X(t)$$

  for all t in a neighborhood of zero. Then there is a unique cdf F_X, whose moments are given by $M_X(t)$, such that $\lim_{n\to\infty} F_{X_n}(u) = F_X(u)$

- convergence of mgfs for $|t| < h$ implies convergence of cdfs (sufficient condition, but not necessary!)

Poisson approximation

- **theorem**: if $X \sim \text{Bin}(n, p)$ and $Y \sim \text{Poisson}(\lambda)$ with $\lambda = np$, then

$$\mathbb{P}(X = x) \;\simeq\; \mathbb{P}(Y = x) \;=\; \frac{e^{-\lambda}\lambda^x}{x!}$$

  for large n and small np

- **proof**: the mgf of Y is

$$M_Y(t) \;=\; \sum_{x=0}^{\infty} e^{tx}\frac{e^{-\lambda}\lambda^x}{x!} \;=\; e^{-\lambda}\sum_{x=1}^{\infty}\frac{(e^t\lambda)^x}{x!} \;=\; e^{-\lambda}e^{e^t\lambda} \;=\; e^{\lambda(e^t-1)}$$

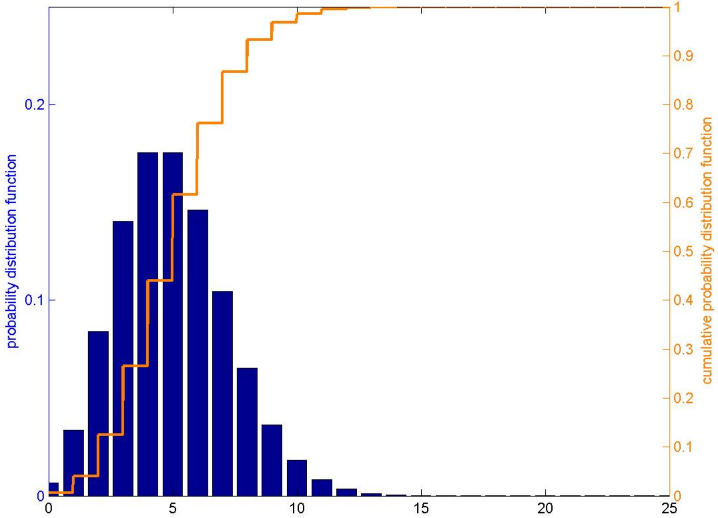
  because $e^x = \sum_{k=0}^{\infty}\frac{x^k}{k!}$. Recall that

$$\begin{aligned} M_X(t) &= [pe^t + (1-p)]^n \;=\; \left[1 + p(e^t - 1)\right]^n \\ &= \left[1 + \frac{np}{n}(e^t - 1)\right]^n \;=\; \left[1 + \frac{\lambda}{n}(e^t - 1)\right]^n \end{aligned}$$

$$\lim_{n\to\infty} M_X(t) \;=\; \lim_{n\to\infty}\left[1 + \frac{\lambda}{n}(e^t - 1)\right]^n \;=\; e^{\lambda(e^t-1)} \;=\; M_Y(t)$$

  since $\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^a$.

looks of the Poisson distribution



mgf of affine transformations

- **theorem**: for any 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**: by definition,

$$
\begin{aligned}
M_{aX+b}(t) &= \mathbb{E}\left[e^{(aX+b)t}\right] & \\
&= \mathbb{E}\left[e^{aXt} e^{bt}\right] & \text{(exponential property)} \\
&= e^{bt}\mathbb{E}\left[e^{(at)X}\right] & (e^{bt} \text{ is constant}) \\
&= e^{bt} M_X(at) & \blacksquare
\end{aligned}
$$

Contents

Leibniz's rule

- interchanging the order of integration and differentiation is common in theoretical statistics, and hence it is convenient to spend some time characterizing conditions under which this operation is legitimate

- fundamental theorem of calculus + chain rule: if $f(x,\theta)$, $a_\theta = a(\theta)$ and $b_\theta = b(\theta)$ are differentiable with respect to θ, then

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{a_\theta}^{b_\theta} f(x,\theta)\,\mathrm{d}x = f(b_\theta,\theta)\,\frac{\mathrm{d}b_\theta}{\mathrm{d}\theta} - f(a_\theta,\theta)\,\frac{\mathrm{d}a_\theta}{\mathrm{d}\theta} + \int_{a_\theta}^{b_\theta} \frac{\partial}{\partial\theta} f(x,\theta)\,\mathrm{d}x$$

  and, in particular,

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{a}^{b} f(x,\theta)\,\mathrm{d}x = \int_{a}^{b} \frac{\partial}{\partial\theta} f(x,\theta)\,\mathrm{d}x$$

  for a and b constants

- Exchanging derivative and integral over a finite range poses no problems.

what happens if range is infinite?

- in principle, the question is really whether we may interchange limits and integration given that a derivative is a special kind of limit

$$\frac{\partial}{\partial \theta} f(x, \theta) = \lim_{\delta \to 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta}$$

  and hence

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial \theta} f(x, \theta) \, \mathrm{d}x = \int_{-\infty}^{\infty} \lim_{\delta \to 0} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} \, \mathrm{d}x,$$

  whereas

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{-\infty}^{\infty} f(x, \theta) \, \mathrm{d}x = \lim_{\delta \to 0} \int_{-\infty}^{\infty} \frac{f(x, \theta + \delta) - f(x, \theta)}{\delta} \, \mathrm{d}x$$

- let's then justify interchanging limits and integration!

Lebesgue's dominated convergence theorem

- **theorem**: suppose the function $h(x, y)$ is continuous at y_0 for each x and that there exists a dominating function $g(x)$ with a finite integral, i.e., $g(x)$ such that

  (i) $|h(x, y)| \leq g(x)$ for all x and y;

  (ii) $\int_{-\infty}^{\infty} g(x)\,\mathrm{d}x < \infty$

  then

$$\lim_{y \to y_0} \int_{-\infty}^{\infty} h(x, y)\,\mathrm{d}x = \int_{-\infty}^{\infty} \lim_{y \to y_0} h(x, y)\,\mathrm{d}x$$

- key condition: the dominating function is well-behaved and hence puts enough discipline on $h(x, y)$ to ensure the validity of interchanging the order of limits and integrals

applying to the difference in the limit

- **theorem**: suppose $f(x,\theta)$ is differentiable at $\theta = \theta_0$, namely

$$\lim_{\delta\to 0} \frac{f(x,\theta_0+\delta) - f(x,\theta_0)}{\delta} = \frac{\partial}{\partial\theta} f(x,\theta)\bigg|_{\theta=\theta_0} \qquad \text{exists for every } x,$$

  and there exists a function $g(x,\theta_0)$ and a constant $\delta_0 > 0$ such that

  (i) $\left|\frac{f(x,\theta_0+\delta)-f(x,\theta_0)}{\delta}\right| \le g(x,\theta_0)$ for all x and $|\delta| \le \delta_0$.

  (ii) $\int_{-\infty}^{\infty} g(x,\theta_0)dx < \infty$

  then

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\int_{-\infty}^{\infty} f(x,\theta)\,\mathrm{d}x\bigg|_{\theta=\theta_0} = \int_{-\infty}^{\infty}\left[\frac{\partial}{\partial\theta} f(x,\theta)\bigg|_{\theta=\theta_0}\right]\mathrm{d}x$$

- the first condition is also known as Lipschitz condition, which imposes smoothness on a function.

Lagrange expansion

- typically, $f(x, \theta)$ is differentiable at all θ, not at just a single value θ_0
- in this case, we may replace the Lipschitz-like condition by another condition that often proves easier to verify by an application of the mean value theorem. It follows that, for fixed x and θ_0, and for $|\delta| \leq \delta_0$
$$\frac{f(x, \theta_0 + \delta) - f(x, \theta_0)}{\delta} = \frac{\partial}{\partial \theta} f(x, \theta)\bigg|_{\theta = \theta_0 + \delta_*(x)}$$
for some $\delta_*(x)$ such that $|\delta_*(x)| \leq \delta_0$
- hence it suffices to find a $g(x, \theta)$ such that $\frac{\partial}{\partial \theta} f(x, \theta)\big|_{\theta = \theta'} \leq g(x, \theta)$ for all θ' such that $|\theta - \theta'| \leq \delta_0$

example: moment recursions

- example: let's calculate $\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathbb{E}(X^k)$ for some integer $k > 0$, with $X \sim \mathrm{Exp}(\lambda)$

- if we could move the differentiation inside the integral, we would have

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathbb{E}(X^k) &= \frac{d}{d\lambda}\int_0^\infty x^k \frac{1}{\lambda} e^{-x/\lambda} dx \\
&= \int_0^\infty \frac{\partial}{\partial\lambda} x^k \frac{1}{\lambda}\, e^{-x/\lambda}\, \mathrm{d}x \\
&= \int_0^\infty \frac{x^k}{\lambda^2}\left(\frac{x}{\lambda} - 1\right) e^{-x/\lambda}\, \mathrm{d}x \\
&= \frac{\mathbb{E}(X^{k+1})}{\lambda^2} - \frac{\mathbb{E}(X^k)}{\lambda}
\end{aligned}
$$

- $\mathbb{E}(X^{k+1}) = \lambda\mathbb{E}(X^k) + \lambda^2 \frac{\mathrm{d}}{\mathrm{d}\lambda}\,\mathbb{E}(X^k)$ recursion makes the calculation of higher-order moments relatively easy, existing for many other distributions

- CB example 2.4.6: if $X \sim N(\mu, 1)$, then $\mathbb{E}(X^{k+1}) = \mu\,\mathbb{E}(X^k) - \frac{\mathrm{d}}{\mathrm{d}\mu}\,\mathbb{E}(X^k)$

example: moment recursions

- to justify the interchange, we bound the derivative of $x^k(1/\lambda)e^{-x/\lambda}$.
- we have that
$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^k e^{-x/\lambda}}{\lambda}\right)\right| = \frac{x^k e^{-x/\lambda}}{\lambda^2}\left|\frac{x}{\lambda}-1\right| \leq \frac{x^k e^{-x/\lambda}}{\lambda^2}\left(\frac{x}{\lambda}+1\right)$$
because $\frac{x}{\lambda} > 0$.
- for some constant δ_0 satisfying $0 < \delta_0 < \lambda$, take
$$g(x,\lambda) = \frac{x^k e^{-x/(\lambda+\delta_0)}}{(\lambda-\delta_0)^2}\left(\frac{x}{\lambda-\delta_0}+1\right)$$
- we then have that
$$\left|\frac{\partial}{\partial\lambda}\left(\frac{x^k e^{-x/\lambda}}{\lambda}\right)\Big|_{\lambda=\lambda'}\right| \leq g(x,\lambda)$$
for all λ' such that $|\lambda'-\lambda| \leq \delta_0$.
- finally, since the exponential distribution has all its moments, $\int_{-\infty}^{\infty} g(x,\lambda)dx < \infty$ as long as $\lambda - \delta_0 > 0$.

interchanging summation and differentiation

- justification for taking the derivative inside the summation is more straightforward than for integration

- **theorem**: if the series $\sum_{x=0}^{\infty} h(x,\theta)$ converges for every $\theta \in (\underline{\theta}, \bar{\theta})$ and

  (a) $\frac{\partial}{\partial\theta} h(x,\theta)$ is continuous in θ for each x

  (b) $\sum_{x=0}^{\infty} \frac{\partial}{\partial\theta} h(x,\theta)$ converges uniformly on every closed bounded subinterval of $(\underline{\theta}, \bar{\theta})$

  then... interchanging summation and differentiation is legitimate

application: mean of the geometric distribution

- example: let $X \sim f_X(x) = \mathbb{P}(X = x) = \theta(1-\theta)^x$, supposing we can interchange derivatives and sums:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta} \sum_{x=0}^{\infty} \theta(1-\theta)^x &= \sum_{x=0}^{\infty} \frac{\mathrm{d}}{\mathrm{d}\theta} \theta(1-\theta)^x \\
&= \sum_{x=0}^{\infty} \left[(1-\theta)^x - x\theta(1-\theta)^{x-1}\right] \\
&= \frac{1}{\theta} \sum_{x=0}^{\infty} \theta(1-\theta)^x - \frac{1}{1-\theta} \sum_{x=0}^{\infty} x\theta(1-\theta)^x
\end{aligned}
$$

- since $\sum_{x=0}^{\infty} \theta(1-\theta)^x = 1$ for all $0 < \theta < 1$, its derivative is 0.
- then

$$
\frac{1}{\theta} = \frac{1}{1-\theta}\mathbb{E}(X) \iff \mathbb{E}(X) = \frac{1-\theta}{\theta}
$$

Contents

Reference:

- Casella and Berger, Ch. 2

Exercises:

- 2.1, 2.4, 2.6, 2.7, 2.9, 2.13-2.18, 2.23-2.28, 2.32-2.33, 2.38.