# Distributions and Inequalities

Ricardo Dahis

PUC-Rio, Department of Economics

Summer 2023

# Contents

# Contents

# Contents

# normal distribution

- plays a central role in statistics: by the Central Limit Theorem, can approximate a large variety of distributions in large sample.

$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < \infty$$

  and we write $X \sim N(\mu, \sigma^2)$.

- property: if $X \sim N(\mu, \sigma^2)$ then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$.

$$\begin{aligned} \mathbb{P}(Z \leq z) &= \mathbb{P}\left(\frac{X-\mu}{\sigma} \leq z\right) \\ &= \mathbb{P}(X \leq z\sigma + \mu) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{z\sigma+\mu} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{t^2}{2}} dt \end{aligned}$$

  defining $t = \frac{x-\mu}{\sigma}$.

## some properties of the normal distribution

- symmetric around $\mu$ (mean/median/mode)
- location-scale distribution
- inflection points at $\mu \pm \sigma$
- Gaussian approximation to a binomial $X \sim \text{Bin}(25, 0.6)$

$$\begin{aligned}
\mathbb{P}(X \leq 13) &= \sum_{x=0}^{13} \binom{25}{x} 0.6^x 0.4^{25-x} = 0.267 \\
\mathbb{P}\left(Z \leq \frac{13 - 25 \times 0.6}{\sqrt{25 \times 0.6 \times 0.4}}\right) &= \mathbb{P}(Z \leq -0.82) = 0.206 \\
\mathbb{P}\left(Z \leq \frac{13.5 - 25 \times 0.6}{\sqrt{25 \times 0.6 \times 0.4}}\right) &= \mathbb{P}(Z \leq -0.61) = 0.271
\end{aligned}$$

# Contents

## exponential family

- pdfs/pmfs that belong to the exponential family are such that

$$f(x|\boldsymbol{\theta}) \;=\; h(x)\, c(\boldsymbol{\theta})\, \exp\left(\sum_{i=1}^{d} w_i(\boldsymbol{\theta}) t_i(x)\right)$$

- $h(x) \geq 0$ and $c(\boldsymbol{\theta}) \geq 0$
- $h(x), t_1(x), \ldots, t_d(x)$ are real-valued functions only of $x$
- $c(\theta), w_1(\boldsymbol{\theta}), \ldots, w_d(\boldsymbol{\theta})$ are real-valued functions only of $\boldsymbol{\theta}$
- examples

| discrete | continuous |
|---|---|
| binomial | beta |
| negative binomial | gamma |
| Poisson | normal |

## binomial exponential family

$$\underline{\text{definition:}} \qquad f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{d} w_i(\boldsymbol{\theta}) t_i(x)\right)$$

- Binomial belongs to the exponential family:

$$\begin{aligned} f(x|p) &= \binom{n}{x} p^x (1-p)^{n-x} = \binom{n}{x} \left(\frac{p}{1-p}\right)^x (1-p)^n \\ &= \binom{n}{x} (1-p)^n \exp\left[x \ln\left(\frac{p}{1-p}\right)\right] \end{aligned}$$

$$\Downarrow$$

$$\begin{aligned} h(x) &= \begin{cases} \binom{n}{x} & \text{if } x = 0, \ldots, n \\ 0 & \text{otherwise} \end{cases} \\ c(p) &= (1-p)^n \\ w_1(p) &= \ln\left(\frac{p}{1-p}\right) \\ t_1(x) &= x \end{aligned}$$

# normal exponential family

definition: $$f(x|\boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left(\sum_{i=1}^{d} w_i(\boldsymbol{\theta}) t_i(x)\right)$$

- The normal distribution also belongs to the exponential family:

$$\begin{aligned}
f(x|\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \\
&= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \cdot \exp\left(-\frac{1}{2\sigma^2}\,x^2 + \frac{\mu}{\sigma^2}\,x\right)
\end{aligned}$$

$$\Downarrow$$

$$\begin{aligned}
h(x) &= 1 \\
c(\mu,\sigma^2) &= \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \\
w_1(\mu,\sigma^2) &= 1/\sigma^2 \;,\quad w_2(\mu,\sigma^2) = \mu/\sigma^2 \\
t_1(x) &= -x^2/2 \;,\quad t_2(x) = x
\end{aligned}$$

## moments of the exponential family

- **theorem** (CB 3.4.2): if $X$ is a random variable with pdf/pmf in the exponential family, then

$$\begin{aligned}
\mathbb{E}\left[\sum_{i=1}^{d} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j}\, t_i(X)\right] &= -\frac{\partial \ln c(\boldsymbol{\theta})}{\partial \theta_j} \\
\operatorname{var}\left[\sum_{i=1}^{d} \frac{\partial w_i(\boldsymbol{\theta})}{\partial \theta_j}\, t_i(X)\right] &= -\frac{\partial^2 \ln c(\boldsymbol{\theta})}{\partial \theta_j^2} - \mathbb{E}\left[\sum_{i=1}^{d} \frac{\partial^2 w_i(\boldsymbol{\theta})}{\partial \theta_j^2}\, t_i(X)\right]
\end{aligned}$$

- main advantage: life is much easier once we replace either integration or summation by differentiation

# binomial mean

- example: binomial mean

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}p}\, w_1(p) &= \frac{\mathrm{d}}{\mathrm{d}p}\ln\left(\frac{p}{1-p}\right) = \frac{1}{p(1-p)} \\
\frac{\mathrm{d}}{\mathrm{d}p}\ln c(p) &= \frac{\mathrm{d}}{\mathrm{d}p}\, n\ln(1-p) = -\frac{n}{1-p} \\
&\Rightarrow \mathbb{E}\left[\frac{X}{p(1-p)}\right] = \frac{n}{1-p} \\
&\Rightarrow \mathbb{E}(X) = np
\end{aligned}$$

- variance identity works in a similar manner

## moments of exponential family proof

- proof: to ensure that the pdf integrates to 1, we have that

$$
\begin{aligned}
c(\theta) &= \left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right) \mathrm{d}x\right]^{-1} \\
\frac{d}{d\theta} \ln c(\theta) &= \frac{d}{d\theta} \ln \left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right) \mathrm{d}x\right]^{-1} \\
&= -\left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right) \mathrm{d}x\right] \cdot \\
&\quad \cdot \left[\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right) \mathrm{d}x\right]^{-2} \cdot \\
&\quad \cdot \frac{d}{d\theta} \int_{-\infty}^{\infty} h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right) \mathrm{d}x \\
&= -c(\theta) \int_{-\infty}^{\infty} \frac{d}{d\theta} \left[h(x) \exp\left(\sum_{i=1}^{d} w_i(\theta) t_i(x)\right)\right] \mathrm{d}x
\end{aligned}
$$

assuming that we can exchange integration and differentiation

## moments of exponential family proof

- proof (cont'd):

$$
\begin{aligned}
&= -c(\theta)\int_{-\infty}^{\infty} h(x)\,\exp\left(\sum_{i=1}^{d} w_i(\theta)t_i(x)\right)\left(\sum_{i=1}^{d} t_i(x)\frac{d}{d\theta}w_i(\theta)\right)\,\mathrm{d}x \\
&= -\mathbb{E}\left[\sum_{i=1}^{d} t_i(x)\frac{d}{d\theta}w_i(\theta)\right]
\end{aligned}
$$

  and so $\mathbb{E}\left[\sum_{i=1}^{d} t_i(x)\frac{d}{d\theta}w_i(\theta)\right] = -\frac{d}{d\theta}\ln c(\theta)$. ∎

- A similar expression holds for the variance.

## keeping track of the support...

- attention to the support: in general, the set of values $x$ for which $f(x|\boldsymbol{\theta}) > 0$ cannot depend on the parameter vector $\boldsymbol{\theta}$ in an exponential family, because otherwise the pdf would not entirely conform to the definition

- example:

$$\begin{aligned} f(x|\theta) &= \frac{1}{\theta} e^{1-\frac{x}{\theta}} \text{ for } 0 < \theta < x < \infty \\ &= \frac{1}{\theta} e^{1-\frac{x}{\theta}} I_{[\theta,\infty]}(x) \end{aligned}$$

  this pdf is not in the exponential family because the indicator function $I_{[\theta,\infty)}(x)$ not only depends both on $x$ and $\theta$, but also cannot be expressed as an exponential

# Contents

# standardizing pdfs

- **theorem** (CB 3.5.1): let $f(x)$ denote a pdf and let $\mu$ and $\sigma > 0$ denote any given constants, then $g(x|\mu,\sigma) = \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ is also a pdf

- **proof**: we must check whether $g(x|\mu,\sigma)$ is a pdf for every value of $\mu$ and $\sigma$ that we may substitute in the formula
  - (i) $f(x) \geq 0$ for all $x$ by definition, and hence $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \geq 0$ as well for all values of $x$, $\mu$ and $\sigma$
  - (ii) $\int_{-\infty}^{\infty} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x = \int_{-\infty}^{\infty} f(y)\, \mathrm{d}y = 1$ with $y = \frac{x-\mu}{\sigma}$ ■

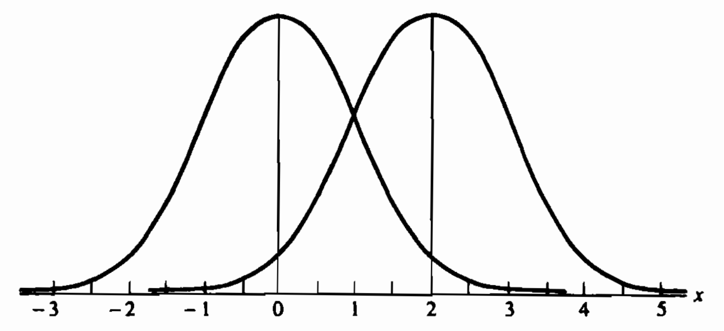
## location family of distributions

- **definition**: let $f(x)$ denote a pdf, then the family of pdfs $f(x-\mu)$, with $-\infty < \mu < \infty$, is called the location family with standard pdf $f(x)$ and with $\mu$ as location parameter

- the location parameter $\mu$ shifts the distribution either to the right (if positive) or to the left (if negative), without altering the shape of the distribution

## members of the same location family

($\mu = 0$ vs $\mu = 2$)



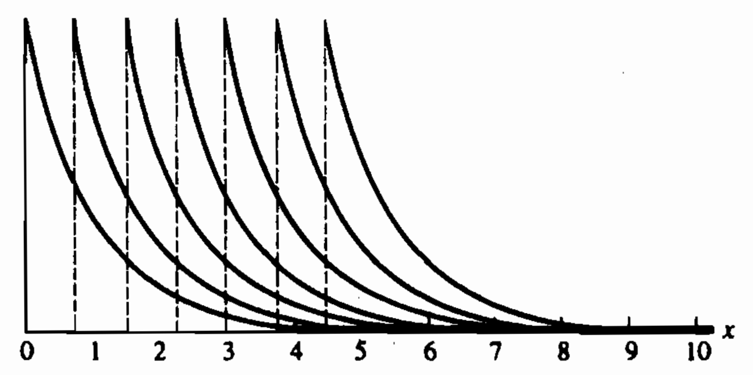
## exponential location family

- it is straightforward to form a location family from $f(x) = e^{-x}$ with $x \geq 0$ by replacing $x$ with $x - \mu$

$$f(x|\mu) = \begin{cases} e^{-(x-\mu)} & \text{if } x - \mu \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

note that $\mu$ now corresponds to a bound on the range of $X$ and hence it is a threshold parameter

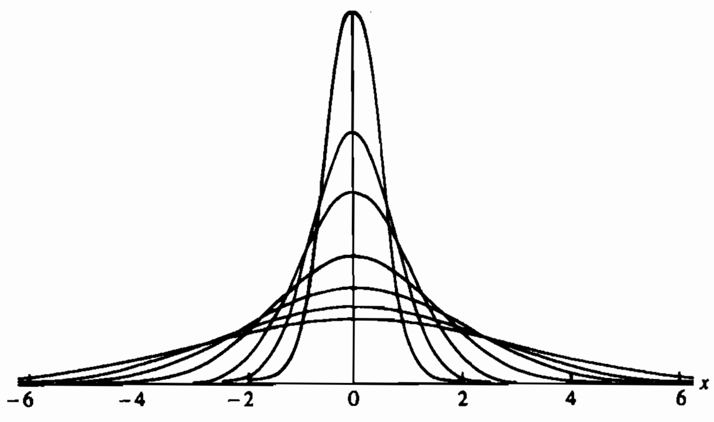
# members of the exponential location family



## introducing the scale parameter...

- **definition**: let $f(x)$ denote a pdf, then the family of pdfs $\frac{1}{\sigma} f(x/\sigma)$ for any $\sigma > 0$ is called the scale family with standard pdf $f(x)$ and scale parameter $\sigma$
- introducing the scale parameter $\sigma$ will either stretch (if $\sigma > 1$) or contract (if $\sigma < 1$) the density, while maintaining the same basic shape
- altogether now: let $f(x)$ denote a pdf, then the family of pdfs $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ for any $-\infty < \mu < \infty$ and $\sigma > 0$ is called the location-scale family with standard pdf $f(x)$, location parameter $\mu$, and scale parameter $\sigma$

## members of the same scale family



## representations within a location-scale family...

- **theorem** (CB 3.5.6): let $f(x)$ denote a pdf, whereas $\mu$ denote a real number and $\sigma$ any positive real number. Then $X$ is a random variable with pdf $\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$ if and only if there exists a random variable $Z$ with pdf $f(z)$ such that $X = \mu + \sigma Z$

- proof: ($\Leftarrow$) Define $g(z) = \mu + \sigma z$. Then $g(\cdot)$ is monotone with $g^{-1} = \frac{x-\mu}{\sigma}$ and $\left|\frac{d}{dx} g^{-1}(x)\right| = \frac{1}{\sigma}$. So the pdf of $X$ is

$$f_X(x) \quad = \quad f_Z\left(g^{-1}(x)\right)\left|\frac{d}{dx} g^{-1}(x)\right| \quad = \quad \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

- proof: ($\Rightarrow$) Define $g(x) = \frac{x-\mu}{\sigma}$, so $g^{-1}(z) = \sigma z + \mu$ with $\left|\frac{d}{dx} g^{-1}(x)\right| = \sigma$. The pdf of $Z$ is

$$f_Z(z) \;=\; f_X\left(g^{-1}(z)\right)\left|\frac{d}{dz} g^{-1}(z)\right| \;=\; \frac{1}{\sigma} f\left(\frac{(\sigma z + \mu) - \mu}{\sigma}\right)\sigma \;=\; f(z)$$

(also, $\sigma Z + \mu = \sigma g(X) + \mu = \sigma\left(\frac{X-\mu}{\sigma}\right) + \mu = X$). ■

# Contents

## Chebychev's inequality

- **theorem** (CB 3.6.1): let $X$ denote a random variable and let $g(x)$ be a nonnegative function, it then follows that

$$\mathbb{P}\big(g(X) \geq r\big) \quad \leq \quad \frac{1}{r}\,\mathbb{E}[g(X)] \quad \text{for any} \quad r > 0$$

- proof:

$$\begin{aligned}
\mathbb{E}[g(X)] &= \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x \\
&\geq \int_{\{x:\, g(x) \geq r\}} g(x) f_X(x)\,\mathrm{d}x \\
&\geq r \int_{\{x:\, g(x) \geq r\}} f_X(x)\,\mathrm{d}x \\
&= r\mathbb{P}\big(g(X) \geq r\big) \qquad \blacksquare
\end{aligned}$$

- very conservative as it applies to **any** distribution!

# Chebychev's inequality

- application: let $g(x) = \frac{(x-\mu)^2}{\sigma^2}$, where $\mu = \mathbb{E}X$, $\sigma^2 = \text{var}X$ and $r = t^2$.
  - By the Chebychev's inequality,

$$\begin{aligned}
\mathbb{P}(g(X) \geq r) &\leq \frac{1}{r}\mathbb{E}[g(X)] \\
\mathbb{P}\left(\frac{(x-\mu)^2}{\sigma^2} \geq t^2\right) &\leq \frac{1}{t^2}\mathbb{E}\left[\frac{(x-\mu)^2}{\sigma^2}\right] = \frac{1}{t^2} \\
&\Downarrow \\
\mathbb{P}(|X-\mu| \geq t\sigma) &\leq \frac{1}{t^2}
\end{aligned}$$

- useful to get universal bounds of $|X-\mu|$. For $t = 2$,

$$\mathbb{P}(|X-\mu| \geq 2\sigma) \leq \frac{1}{2^2} = .25$$

that is, there at least a 75% chance that a random variable is within $2\sigma$ of its mean (regardless of the distribution of $X$!)

## normal probability inequality

- **theorem** (CB 3.6.3): if $Z \sim N(0,1)$, then

$$\mathbb{P}(|Z| \geq t) \;\leq\; \sqrt{\frac{2}{\pi}}\,\frac{e^{-t^2/2}}{t} \qquad \text{for all } t > 0$$

- proof:

$$\begin{aligned}
\mathbb{P}(Z \geq t) &= \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2}\,\mathrm{d}x \\
&\leq \frac{1}{\sqrt{2\pi}} \int_t^\infty \frac{x}{t}\, e^{-x^2/2}\,\mathrm{d}x && \text{given } x > t > 0 \\
&= \frac{1}{\sqrt{2\pi}}\,\frac{e^{-t^2/2}}{t},
\end{aligned}$$

yielding the result as $\mathbb{P}(|Z| \geq t) = 2\mathbb{P}(Z \geq t)$ ■

- Vast improvement over Chebychev: $\sqrt{(2/\pi)}e^{-2}/2 = .054$

## Chebychev or Markov

- **warning**: there are many versions of these theorems.
- Some authors refer to the Markov inequality

$$\mathbb{P}\big(X \geq r\big) \quad \leq \quad \frac{1}{r}\mathbb{E}(X) \quad \text{for any} \quad r > 0$$

- and to the Chebychev inequality as

$$\begin{array}{rcl} \mathbb{P}\big(|X-\mu| \geq t\big) & \leq & \dfrac{1}{t^2}\,\mathrm{var}\,X \quad \text{for any} \quad t > 0 \\ & \Uparrow & \\ \mathbb{P}\big((X-\mu)^2 \geq t^2\big) & \leq & \dfrac{1}{t^2}\,\mathrm{var}\,X \quad \text{for any} \quad t > 0 \end{array}$$

# Stein's lemma

- there is a wide array of identities that rely on integration by parts, of which the first is due to Charles Stein
- **theorem** (CB 3.6.5): let $X \sim N(\mu, \sigma^2)$ and let $g$ denote a differentiable function such that $\mathbb{E}|g'(X)| < \infty$, then $\mathbb{E}\big[g(X)(X - \mu)\big] = \sigma^2\, \mathbb{E}[g'(X)]$

## Stein's lemma

- proof:

$$\mathbb{E}[g(X)(X-\mu)] \quad = \quad \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} g(x)(x-\mu) e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x$$

- Refresher on integral by parts: $\int_a^b u(x)v'(x)dx = [u(x)v(x)]_a^b - \int_a^b u'(x)v(x)dx$.
- Using integration by parts and setting

$$\begin{aligned} u &= g(x) &\Rightarrow \quad \mathrm{d}u &= g'(x)\,\mathrm{d}x \\ \mathrm{d}v &= (x-\mu)e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\mathrm{d}x &\Rightarrow \quad v &= -\sigma^2 e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \end{aligned}$$

in the LHS yields

$$= \quad \frac{1}{\sqrt{2\pi}} \left[-\sigma^2 g(x) e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}\right]_{-\infty}^{\infty} + \sigma^2\,\mathbb{E}[g'(X)],$$

whereas $\mathbb{E}|g'(X)| < \infty$ ensures that the first term is zero. ■

## higher-order moments of a normal distribution

- Stein's lemma is very useful to compute the higher-order moments of a normal distribution

$$\begin{aligned}
\mathbb{E}(X^3) &= \mathbb{E}\big[X^2(X-\mu+\mu)\big] \\
&= \mathbb{E}\big[X^2(X-\mu)\big] + \mu\,\mathbb{E}(X^2) \\
&= 2\sigma^2\mathbb{E}(X) + \mu(\sigma^2+\mu^2) \\
&= 3\mu\sigma^2 + \mu^3 \\
&\Rightarrow \mathbb{E}(Z^3) = 0 \qquad \text{if } Z = (X-\mu)/\sigma
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}(X^4) &= \mathbb{E}\big[X^3(X-\mu+\mu)\big] \\
&= \mathbb{E}\big[X^3(X-\mu)\big] + \mu\,\mathbb{E}(X^3) \\
&= 3\sigma^2\mathbb{E}(X^2) + 3\mu^2\sigma^2 + \mu^4 \\
&= 3\sigma^2(\sigma^2+\mu^2) + 3\mu^2\sigma^2 + \mu^4 \\
&= 3(\sigma^4 + 2\sigma^2\mu^2 + \mu^2) + \mu^4 \\
&= 3(\sigma^2+\mu)^2 + \mu^4 \\
&\Rightarrow \mathbb{E}(Z^4) = 3 \qquad \text{if } Z = (X-\mu)/\sigma
\end{aligned}$$

# Jensen's Inequality

- **theorem**: (Jensen's Inequality) Let $g : \mathbb{R} \to \mathbb{R}$ convex. Then $g(\mathbb{E}X) \leq \mathbb{E}g(X)$.
- proof: since $g$ is convex, there exists a linear function $l : \mathbb{R} \to \mathbb{R}$ such that $l \leq g$ and $l(\mathbb{E}X) = g(\mathbb{E}X)$. It follows that

$$\begin{aligned} \mathbb{E}g(X) &\geq \mathbb{E}l(x) \\ &= l(\mathbb{E}X) \\ &= g(\mathbb{E}X) \end{aligned}$$

$\blacksquare$

# Contents

Reference:

- Casella and Berger, Ch. 3

Exercises:

- 3.1–3.3, 3.5–3.9, 3.12–3.15, 3.17, 3.20, 3.23–3.26, 3.30–3.32, 3.37–3.39.