Statistics for Data Science: Week 2

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

rajita.chandak@epfl.ch, myrto.limnios@epfl.ch

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X, Y \text{ are discrete,} \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

[1] measurable

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X, Y \text{ are discrete,} \\ \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

- Precisely the expectation of the conditional distribution.

[1] measurable

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X, Y \text{ are discrete,} \\ \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

- Precisely the expectation of the conditional distribution.
- Note that $\mathbb{E}[X|Y=y] = q(y)$ results in a function of only y.

[1] measurable

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X, Y \text{ are discrete,} \\ \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X, Y \text{ are continuous.} \end{cases}$$

- Precisely the expectation of the conditional distribution.
- Note that $\mathbb{E}[X|Y=y] = q(y)$ results in a function of only y.
- One can plug Y into $q(\cdot)$ and consider $Z = \underline{q(Y)}$ as a random variable itself.

[1] measurable

We can calculate the conditional expectation of a random variable X given that another random variable Y took the value y as

$$\mathbb{E}[X|Y=y] = \begin{cases} \sum_{x\in\mathcal{X}} x\,\mathbb{P}[X=x|Y=y], & \text{if } X,Y \text{ are discrete,} \\ \int_{-\infty}^{+\infty} x\, f_{X|Y}(x|y)dx, & \text{if } X,Y \text{ are continuous.} \end{cases}$$

- Precisely the expectation of the conditional distribution.
- Note that $\mathbb{E}[X|Y=y] = q(y)$ results in a function of only y.
- One can plug Y into $q(\cdot)$ and consider $Z = q(Y)$ as a random variable itself.
- Important property/interpretation:

MSE

$$\mathbb{E}[X|Y] = \underline{\arg\min_g \mathbb{E}\,\|X - \underline{g(Y)}\|^2}$$

Among all functions[1] of Y, $\mathbb{E}[X|Y]$ best approximates X in mean square.

[1] measurable

$$(a+b)^2 = a^2+b^2+2ab$$

$$\underset{g \in \mathcal{G}}{\arg\min}\ \mathbb{E}[\|X - g(X)\|^2]$$

$$= \mathbb{E}[\| \underbrace{X - \mathbb{E}[Y|X]}_{a} + \underbrace{\mathbb{E}[Y|X] - g(X)}_{b} \|^2]$$

$$= \mathbb{E}[\|X - \mathbb{E}[Y|X]\|^2] + \mathbb{E}[\|\mathbb{E}[Y|X] - g(X)\|^2]$$

$$+ 2\,\mathbb{E}[(X - \mathbb{E}[Y|X])\underbrace{(\mathbb{E}[Y|X] - g(X))}_{=0}]$$

$$\min \mathbb{E}[\| \underbrace{\mathbb{E}[Y|X] - g(X)}_{\geq 0} \|^2]$$

$$g(X) = \mathbb{E}[Y|X]$$

Important properties of $\mathbb{E}[X|Y]$:

Important properties of $\mathbb{E}[X|Y]$:

1. 'Tower property': $\mathbb{E}_Y\Big[\underbrace{\mathbb{E}[X|Y]}_{g(Y)}\Big] = \mathbb{E}[X]$

Important properties of $\mathbb{E}[X|Y]$:

1. 'Tower property': $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$

2. If X independent of Y, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

$\mathbb{P}(X|Y) = \mathbb{P}(X)$

$\int x f_{X|Y} = \int x f_X = \mathbb{E}[X]$

Important properties of $\mathbb{E}[X|Y]$:

1. 'Tower property': $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$

   $\mathbb{E}[aX] = a\mathbb{E}[X]$

2. If X independent of Y, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.

3. $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$ (taking out known factors)

   $Y = y \rightarrow g(Y) = g(y)$

Important properties of $\mathbb{E}[X|Y]$:

1. 'Tower property': $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$
2. If X independent of Y, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
3. $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$ (taking out known factors)
4. Linearity: $\mathbb{E}[aX_1 + X_2|Y] = a\mathbb{E}[X_1|Y] + \mathbb{E}[X_2|Y]$.

Important properties of $\mathbb{E}[X|Y]$:

1. 'Tower property': $\mathbb{E}\Big[\mathbb{E}[X|Y]\Big] = \mathbb{E}[X]$
2. If X independent of Y, then $\mathbb{E}[X|Y] = \mathbb{E}[X]$.
3. $\mathbb{E}[g(Y)X|Y] = g(Y)\mathbb{E}[X|Y]$ (taking out known factors)
4. Linearity: $\mathbb{E}[aX_1 + X_2|Y] = a\mathbb{E}[X_1|Y] + \mathbb{E}[X_2|Y]$.
5. Monotonicity: $\underline{X_1 \leq X_2} \implies \mathbb{E}[X_1|Y] \leq \mathbb{E}[X_2|Y]$, $\mathbb{E}[X_1] \leq \mathbb{E}[X_2]$

The conditional variance of X given Y is defined as

$$\mathrm{var}[X|Y] = \mathbb{E}\Big[\,(X - \mathbb{E}[X|Y])^2\,\Big|\,Y\Big] = \mathbb{E}[X^2|Y] - \big(\mathbb{E}[X|Y]\big)^2$$

The conditional variance of X given Y is defined as

$$\mathrm{var}[X|Y] = \mathbb{E}\Big[(X - \mathbb{E}[X|Y])^2 \,\Big|\, Y\Big] = \mathbb{E}[X^2|Y] - \big(\mathbb{E}[X|Y]\big)^2$$

The law of total variance states that

$$\mathrm{var}(X) = \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left(\mathbb{E}[X|Y]\right)$$

The conditional variance of X given Y is defined as

$$\mathrm{var}[X|Y] = \mathbb{E}\Big[(X - \mathbb{E}[X|Y])^2 \Big| Y\Big] = \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2$$

The law of total variance states that

$$\mathrm{var}(X) = \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left(\mathbb{E}[X|Y]\right)$$

Proof:

$$\begin{aligned}
\mathrm{var}(X) &\overset{\text{def}}{=} \mathbb{E}[X^2] - \mathbb{E}^2[X] \\
&\overset{\text{tower property}}{=} \mathbb{E}\left[\mathbb{E}[X^2|Y]\right] - \mathbb{E}^2\left[\mathbb{E}[X|Y]\right] \\
&= \mathbb{E}\Big[\mathrm{var}[X|Y] + \mathbb{E}^2[X|Y]\Big] - \mathbb{E}^2\left[\mathbb{E}[X|Y]\right] \\
&\overset{\text{linearity of } \mathbb{E}}{=} \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathbb{E}\left[\mathbb{E}^2[X|Y]\right] - \mathbb{E}^2\left[\mathbb{E}[X|Y]\right] \\
&= \mathbb{E}\left[\mathrm{var}[X|Y]\right] + \mathrm{var}\left(\mathbb{E}[X|Y]\right). \quad \text{" def of var.}
\end{aligned}$$

matrix

The covariance matrix of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$, say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$, is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \underbrace{\mathrm{cov}(Y_i, Y_j)}_{\mathrm{cov}(X,Y)} = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \le i \le j \le d.$$

$\mathrm{cov}(Y_i, Y_j) = \mathrm{cov}(Y_j, Y_i)$

$$\begin{bmatrix} \mathrm{Var}(Y_1) & \mathrm{cov}(Y_1, Y_2) & \cdots & \mathrm{cov}(Y_1, Y_d) \\ \vdots & \ddots & & \\ \vdots & & \ddots & \\ \vdots & & & \mathrm{Var}(Y_d) \end{bmatrix}$$

The covariance matrix of a random vector $\boldsymbol{Y} = (Y_1, \ldots, Y_d)^\top$, say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$, is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \mathrm{cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \leq i \leq j \leq d.$$

That is, the covariance matrix encodes the variances of the coordinates of Y (on the diagonal) and the pairwise covariances between any two coordinates of $\boldsymbol{Y}$ (off the diagonal).

The covariance matrix of a random vector $\boldsymbol{Y} = (Y_1, \dots, Y_d)^\top$, say $\boldsymbol{\Omega} = \{\Omega_{ij}\}$, is a $d \times d$ symmetric matrix with entries

$$\Omega_{ij} = \operatorname{cov}(Y_i, Y_j) = \mathbb{E}[(Y_i - \mathbb{E}[Y_i])(Y_j - \mathbb{E}[Y_j])], \quad 1 \le i \le j \le d.$$

That is, the covariance matrix encodes the variances of the coordinates of Y (on the diagonal) and the pairwise covariances between any two coordinates of $\boldsymbol{Y}$ (off the diagonal).

If we write

$$\boldsymbol{\mu} = \underline{\mathbb{E}[\boldsymbol{Y}]} = \underline{(\mathbb{E}[Y_1], \dots, \mathbb{E}[Y_d])}^\top \qquad \begin{bmatrix} Y_1 \\ \vdots \\ Y_d \end{bmatrix}$$

for the mean vector of $\boldsymbol{Y}$, then

$$\mathbb{E}[\underline{(\boldsymbol{Y} - \boldsymbol{\mu})^{T}(\boldsymbol{Y} - \boldsymbol{\mu})^{\cancel{\top}}}] = \mathbb{E}[\underline{\boldsymbol{Y}^{T}\boldsymbol{Y}^{\cancel{\top}}}] - \boldsymbol{\mu}^{T}\boldsymbol{\mu}^{\cancel{\top}}.$$

Similarly to the vector case, the expectation of a matrix with random entries is the matrix of expectations of the random entries.

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$. Σ

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- PSD: for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$. Positive semi-definite

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- PSD: for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of $\boldsymbol{AY}$ are $\boldsymbol{A\mu}$ and $\boldsymbol{A\Omega A}^\top$, respectively.

Cov(AY, AY)

$\mathbb{E}[(X-\mathbb{E}[X])^2] \geq 0$

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- PSD: for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of $\boldsymbol{AY}$ are $\boldsymbol{A\mu}$ and $\boldsymbol{A\Omega A}^\top$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, the variance of $\underline{\boldsymbol{\beta}^\top \boldsymbol{Y}}$ is $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$. ≥ 0

Let $\boldsymbol{Y}$ be a random $d \times 1$ with mean vector $\boldsymbol{\mu}$ be the mean vector and covariance matrix $\boldsymbol{\Omega}$.

- PSD: for any $\boldsymbol{\beta} \in \mathbb{R}^d$, we have $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta} \geq 0$.
- If $\boldsymbol{A}$ is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of $\boldsymbol{AY}$ are $\boldsymbol{A\mu}$ and $\boldsymbol{A\Omega A}^\top$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, the variance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ is $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.
- If $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ are deterministic vectors, the covariance of $\boldsymbol{\beta}^\top \boldsymbol{Y}$ with $\boldsymbol{\gamma}^\top \boldsymbol{Y}$ is $\boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.

$\mathrm{cov}(\beta^T Y, \gamma^T Y) = \gamma^T \Omega \beta$

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

$X \geq 0$

Proof:

$\mathbb{E}[X] = \mathbb{E}[X \, \mathbb{1}\{X \geq \varepsilon\} + X \, \mathbb{1}\{X < \varepsilon\}]$

$\mathbb{1}\{X \geq \varepsilon\} = \begin{cases} 1 & \text{if } X \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}$

$= \mathbb{E}[X \, \mathbb{1}\{X \geq \varepsilon\}] + \mathbb{E}[X \, \mathbb{1}\{ < \varepsilon\}]$

$\mathbb{E}[\mathbb{1}\{X \geq \varepsilon\}] = \mathbb{P}(X \geq \varepsilon)$

$\int_{\varepsilon}^{\infty} 1 \, dF_X = \mathbb{P}(X \geq \varepsilon)$

$= f(x)dx$

$= \underbrace{\mathbb{P}(X \geq \varepsilon)}_{\mathbb{E}[\mathbb{1}\{\cdot\}]} \cdot \mathbb{E}[X \mid X \geq \varepsilon] + \mathbb{P}(X < \varepsilon)\, \mathbb{E}[X \mid X < \varepsilon]$

$\geq \varepsilon$

≥ 0

$\geq \boxed{\varepsilon \, \mathbb{P}(X \geq \varepsilon)}$

---

[2] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0,1]$.

$$\mathbb{E}[X] \geq \varepsilon \, \mathbb{P}(X \geq \varepsilon)$$

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}$$

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

Proof

$$Y = |X - \mathbb{E}[X]| \geq 0$$

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq \varepsilon^2)$$

$$\leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{\varepsilon^2}$$

$$= \frac{\mathrm{Var}(X)}{\varepsilon^2}$$

[2] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

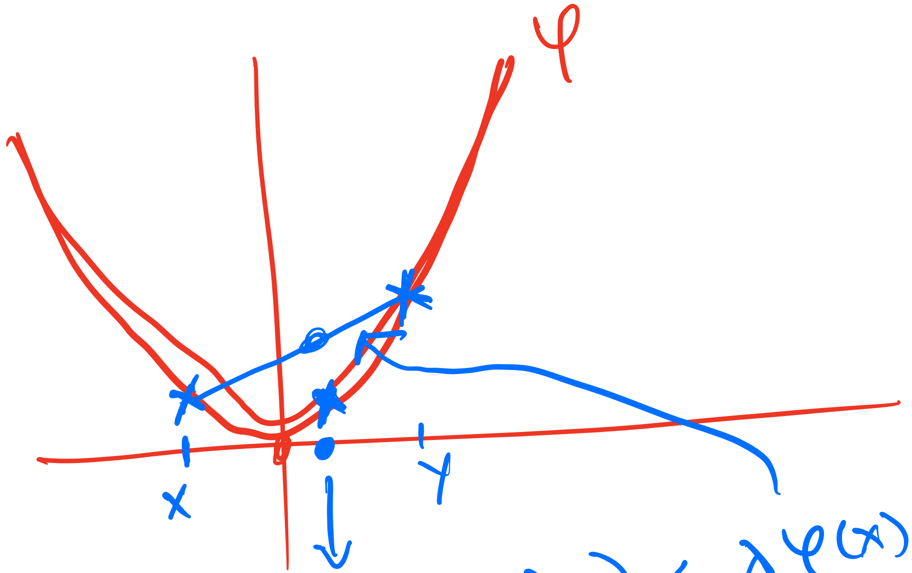
Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\mathrm{var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

For any convex[2] function $\varphi : \mathbb{R} \to \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then one has

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen]}$$

[2] Recall that a function φ is convex if $\varphi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.



$$\varphi(\lambda x+(1-\lambda)y) \leq \underbrace{\lambda\varphi(x)+(1-\lambda)\varphi(y)}$$

Given X be a non-negative random variable. Then, given any $\epsilon > 0$,

$$\mathbb{P}[X \geq \epsilon] \leq \frac{\mathbb{E}[X]}{\epsilon} \qquad \text{[Markov]}$$

Let X be a random variable with finite mean $\mathbb{E}[X] < \infty$. Then, given any $\epsilon > 0$,

$$\mathbb{P}\Big[|X - \mathbb{E}[X]| \geq \epsilon\Big] \leq \frac{\text{var}[X]}{\epsilon^2} \qquad \text{[Chebyschev]}$$

For any convex[2] function $\varphi : \mathbb{R} \to \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then one has

$$\varphi\Big(\mathbb{E}[X]\Big) \leq \mathbb{E}[\varphi(X)] \qquad \text{[Jensen]}$$

Let X be a real random variable with $\mathbb{E}[X^2] < \infty$. Let $g : \mathbb{R} \to \mathbb{R}$ be a non decreasing function such that $\mathbb{E}[g^2(X)] < \infty$. Then,

$x \uparrow \Rightarrow g(x) \downarrow$

$$\text{cov}[X, g(X)] \geq 0 \qquad \text{[Monotonicity and Covariance]}$$

[2] Recall that a function φ is convex if $\varphi(\lambda x + (1 - \lambda)y) \leq \lambda\varphi(x) + (1 - \lambda)\varphi(y)$ for all x, y, and $\lambda \in [0, 1]$.

Let X be a random variable taking values in $\mathbb{R}$. The moment generating function (MGF) of X is defined as

$$M_X(t) : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$$

$$M_X(t) = \mathbb{E}\left[e^{tX}\right], \qquad t \in \mathbb{R}.$$

When $M_X(t), M_Y(t)$ exist (are finite) for $t \in I \ni 0$, then:

- $\mathbb{E}[|X|^k] < \infty$ and $\mathbb{E}[X^k] = \frac{d^k M_X}{dt^k}(0)$, for all $k \in \mathbb{N}$.
- $M_X = M_Y$ on I if and only if $F_X = F_Y$
- $M_{X+Y} = M_X M_Y$ when X and Y are independent

Similarly, for a random vector $\boldsymbol{X}$ in $\mathbb{R}^d$, the MGF is

$$M_{\boldsymbol{X}}(\boldsymbol{u}) : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$$

$$M_{\boldsymbol{X}}(\boldsymbol{u}) = \mathbb{E}\left[e^{\boldsymbol{u}^\top \boldsymbol{X}}\right], \qquad \boldsymbol{u} \in \mathbb{R}^d.$$

and has analogous properties.

$$e^{tx} = 1 + tx + \frac{t^2 x^2}{2!} + \frac{t^3 x^3}{3!} + \cdots$$

$$\mathbb{E}[e^{tx}] = \mathbb{E}[1] + \mathbb{E}[tx] + \frac{\mathbb{E}[t^2 x^2]}{2!} + \cdots$$

$$\mathbb{E}[x^m] \Rightarrow \left.\frac{\partial^m \mathbb{E}[e^{tx}]}{\partial t^m}\right|_{t=0}$$

$$\left.\frac{\partial\, \mathbb{E}[e^{tx}]}{\partial t}\right|_{t=0} = \mathbb{E}[x] + \mathbb{E}\left[\frac{2tx^2}{2!}\right] + \cdots$$

$$= \mathbb{E}[x]$$

Elementary Distributions Factsheet

A random variable X is said to follow the Bernoulli distribution with parameter $p \in (0, 1)$, denoted $X \sim \text{Bern}(p)$, if

1. $\mathcal{X} = \{0, 1\}$, $= \Omega$
2. $f(x; p) = p\mathbf{1}\{x = 1\} + (1 - p)\mathbf{1}\{x = 0\}$.

$\mathbb{P}(X=1)$

The mean, variance and moment generating function of $X \sim \text{Bern}(p)$ are given by

$$\mathbb{E}[X] = p, \qquad \text{var}[X] = p(1 - p), \qquad \boxed{M(t) = 1 - p + pe^t.}$$

$\mathbb{E}[X] = \mathbb{E}[p\mathbb{1}\{X=1\} + (1-p)\mathbb{1}\{X=0\}]$

$= 1 \cdot p + 0 \cdot (1-p) = p$

A random variable X is said to follow the Binomial distribution with parameters $p \in (0, 1)$ and $n \in \mathbb{N}$, denoted $X \sim \text{Binom}(n, p)$, if

1. $\mathcal{X} = \{0, 1, 2, ..., n\}$,
2. $f(x; n, p) = \binom{n}{x} p^x (1 - p)^{n-x}$.

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

The mean, variance and moment generating function of $X \sim \text{Binom}(n, p)$ are given by

$$\mathbb{E}[X] = np, \qquad \text{var}[X] = np(1 - p), \qquad M(t) = (1 - p + pe^t)^n.$$

A random variable X is said to follow the Binomial distribution with parameters $p \in (0,1)$ and $n \in \mathbb{N}$, denoted $X \sim \text{Binom}(n,p)$, if

1. $\mathcal{X} = \{0, 1, 2, ..., n\}$,
2. $f(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$.

The mean, variance and moment generating function of $X \sim \text{Binom}(n,p)$ are given by

$$\mathbb{E}[X] = np, \qquad \mathrm{var}[X] = np(1-p), \qquad M(t) = (1 - p + pe^t)^n.$$

- If $X = \sum_{i=1}^n Y_i$ where $Y_i \overset{iid}{\sim} Bern(p)$, then $X \sim \text{Binom}(n,p)$.

A random variable X is said to follow the Geometric distribution with parameter $p \in (0,1)$, denoted $X \sim \mathrm{Geom}(p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p) = (1-p)^x p$.

The mean, variance and moment generating function of $X \sim \mathrm{Geom}(p)$ are given by

$$\mathbb{E}[X] = \frac{1-p}{p}, \qquad \mathrm{var}[X] = \frac{(1-p)}{p^2}, \qquad M(t) = \frac{p}{1-(1-p)e^t}$$

< 0

the latter for $t < -\log(1-p)$.

- Let $\{Y_i\}_{i \geq 1}$ be an infinite collection of random variables, where $Y_i \overset{iid}{\sim} \mathit{Bern}(p)$. Let $T = \min\{k \in \mathbb{N} : Y_k = 1\} - 1$. Then $T \sim \mathit{Geom}(p)$.

A random variable X is said to follow the Negative Binomial distribution with parameters $p \in (0,1)$ and $r > 0$, denoted $X \sim \text{NegBin}(r, p)$, if

Bin (n, p)

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x; p, r) = \binom{x+r-1}{x}(1-p)^x p^r$.

The mean, variance and moment generating function of $X \sim \text{NegBin}(r, p)$ are given by

$$\mathbb{E}[X] = r\frac{1-p}{p}, \qquad \text{var}[X] = r\frac{(1-p)}{p^2}, \qquad M(t) = \frac{p^r}{[1-(1-p)e^t]^r},$$

the latter for $t < -\log(1-p)$.

A random variable X is said to follow the Negative Binomial distribution with parameters $p \in (0,1)$ and $r > 0$, denoted $X \sim \text{NegBin}(r,p)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x;p,r) = \binom{x+r-1}{x}(1-p)^x p^r$.

The mean, variance and moment generating function of $X \sim \text{NegBin}(r,p)$ are given by

$$\mathbb{E}[X] = r\frac{1-p}{p}, \qquad \mathrm{var}[X] = r\frac{(1-p)}{p^2}, \qquad M(t) = \frac{p^r}{[1-(1-p)e^t]^r},$$

the latter for $t < -\log(1-p)$.

- If $X = \sum_{i=1}^r Y_i$ where $Y_i \overset{iid}{\sim} Geom(p)$, then $X \sim \text{NegBin}(r,p)$.

A random variable X is said to follow the Poisson distribution with parameters $\lambda > 0$, denoted $X \sim \text{Poisson}(\lambda)$, if

1. $\mathcal{X} = \{0\} \cup \mathbb{N}$,
2. $f(x; \lambda) = e^{-\lambda} \dfrac{\lambda^x}{x!}$.

The mean, variance and moment generating function of $X \sim \text{Poisson}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda, \qquad \text{var}[X] = \lambda, \qquad M(t) = \exp\{\lambda(e^t - 1)\}.$$

- Let $\{X_n\}_{n\geq 1}$ be a sequence of $Binom(n, p_n)$ random variables, such that $p_n = \lambda/n$, for some constant $\lambda > 0$. Then $f_{X_n} \overset{n\to\infty}{\longrightarrow} f_Y$, where $Y \sim Poisson(\lambda)$.
- Let $X \sim Poisson(\lambda)$ and $Y \sim Poisson(\mu)$ be independent. The conditional distribution of X given $X + Y = k$ is $Binom(k, \lambda/(\lambda + \mu))$ (useful in contingency tables).

$F_{X|X+Y=k}$

sex \ left/right handed	L	R
F	P(F, L)	P(F, R)
M	P(M, L)	P(M, R)

A random vector $\boldsymbol{X}$ in $\mathbb{R}^k$ said to follow the Multinomial distribution with parameters $n \in \mathbb{N}$ and $p = (p_1, \ldots, p_k) \in (0,1)^k$, such that $\sum_{i=1}^k p_i = 1$, denoted $\boldsymbol{X} \sim \text{Multi}(n; p_1, \ldots, p_k)$, if

1. the sample space is $\{0, 1, \ldots, n\}^k$, and
2. $f(x_1, \ldots, x_k; n, \{p_i\}_{i=1}^k) = \dfrac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \mathbf{1}\left\{\sum_{i=1}^k x_i = n\right\}.$

The mean, variance covariance and moment generating function are

$$\mathbb{E}[X_i] = np_i, \qquad \text{Var}[X_i] = np_i(1 - p_i), \qquad \text{cov}(X_i, X_j) = -np_i p_j$$

$$M(u_1, \ldots, u_k) = \left(\sum_{i=1}^k p_i e^{u_i}\right)^n .$$

Generalises binomial: n independent trials, with k possible outcomes.

A random vector $\boldsymbol{X}$ in $\mathbb{R}^k$ said to follow the Multinomial distribution with parameters $n \in \mathbb{N}$ and $p = (p_1, \ldots, p_k) \in (0,1)^k$, such that $\sum_{i=1}^k p_i = 1$, denoted $\boldsymbol{X} \sim \text{Multi}(n; p_1, \ldots, p_k)$, if

1. the sample space is $\{0, 1, \ldots, n\}^k$, and
2. $f(x_1, \ldots, x_k; n, \{p_i\}_{i=1}^k) = \frac{n!}{x_1! \ldots x_k!} p_1^{x_1} \ldots p_k^{x_k} \mathbf{1}\left\{\sum_{i=1}^k x_i = n\right\}$.

The mean, variance covariance and moment generating function are

$$\mathbb{E}[X_i] = np_i, \qquad \mathrm{Var}[X_i] = np_i(1 - p_i), \qquad \mathrm{cov}(X_i, X_j) = -np_ip_j$$

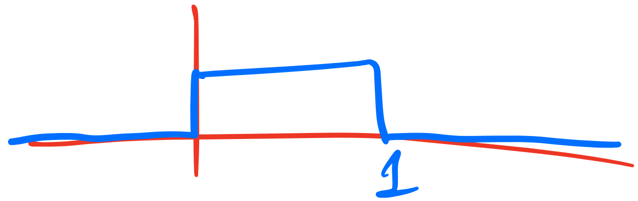
$$M(u_1, \ldots, u_k) = \left(\sum_{i=1}^k p_i e^{u_i}\right)^n.$$

Generalises binomial: n independent trials, with k possible outcomes.

Lemma (Poisson and Multinomial)

If $X_i \sim Poiss(\lambda_i)$, $i = 1, \ldots, k$ are independent, then the conditional distribution of $\boldsymbol{X} = (X_1, \ldots, X_k)^\top$ given $\sum_{i=1}^k X_i = n$ is $\text{Multi}(n; p_1, \ldots, p_k)$, with

$$p_i = \frac{\lambda_i}{\lambda_1 + \ldots + \lambda_k}.$$



A random variable X is said to follow the uniform distribution with parameters $-\infty < \theta_1 < \theta_2 < \infty$, denoted $X \sim \text{Unif}(\theta_1, \theta_2)$, if

$$f_X(x;\theta) = \begin{cases} (\theta_2 - \theta_1)^{-1} & \text{if } x \in (\theta_1, \theta_2), \\ 0 & \text{otherwise.} \end{cases}$$

$\frac{1}{\theta_2 - \theta_1}$

The mean, variance and moment generating function of $X \sim \text{Unif}(\theta_1, \theta_2)$ are given by

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \qquad \text{var}[X] = (\theta_2 - \theta_1)^2/12, \qquad M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}, \ t \neq 0$$

$$M(0) = 1.$$

$\mathbb{E}[e^{tX}] = 1 + \ldots$

A random variable X is said to follow the exponential distribution with parameter $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if

ρ or λ

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

The mean, variance and moment generating function of $X \sim \text{Exp}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \text{var}[X] = \lambda^{-2}, \qquad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

If X, Y are independent exponential random variables with rates λ_1 and λ_2, then $Z = \min\{X, Y\}$ is also exponential with rate $\lambda_1 + \lambda_2$.

$$f_Z = (\lambda_1 + \lambda_2) e^{-(\lambda_1+\lambda_2) z}$$

A random variable X is said to follow the exponential distribution with parameter $\lambda > 0$, denoted $X \sim \text{Exp}(\lambda)$, if

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

The mean, variance and moment generating function of $X \sim \text{Exp}(\lambda)$ are given by

$$\mathbb{E}[X] = \lambda^{-1}, \qquad \mathrm{var}[X] = \lambda^{-2}, \qquad M(t) = \frac{\lambda}{\lambda - t}, \quad t < \lambda.$$

If X, Y are independent exponential random variables with rates λ_1 and λ_2, then $Z = \min\{X, Y\}$ is also exponential with rate $\lambda_1 + \lambda_2$.

Memorylessness property:

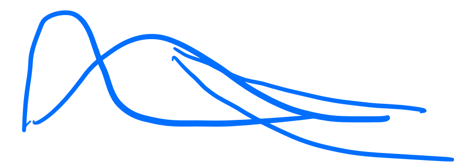
1. Let $X \sim Exp(\lambda)$. Then $\mathbb{P}[X \geq x + t | X \geq t] = \mathbb{P}[X \geq x]$.
2. Conversely: if X is a random variable such that $\mathbb{P}(X > 0) > 0$ and

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \qquad \forall t, s \geq 0,$$

then there exits a $\lambda > 0$ such that $\boxed{X \sim Exp(\lambda)}$.

A random variable X is said to follow the gamma distribution with parameters $r > 0$ and $\lambda > 0$ (the *shape* and *rate* parameters, respectively), denoted $X \sim \text{Gamma}(r, \lambda)$, if

α, β



$$f_X(x; r, \lambda) = \begin{cases} \frac{\lambda^r}{\Gamma(r)} x^{r-1} e^{-\lambda x}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \text{Gamma}(r, \lambda)$ are given by

$$\mathbb{E}[X] = r/\lambda, \qquad \text{var}[X] = r/\lambda^2, \qquad M(t) = \left(\frac{\lambda}{\lambda - t}\right)^r, \quad t < \lambda.$$

- If $Y_1, \ldots, Y_r \overset{iid}{\sim} Exp(\lambda)$, then $Y = \sum_{i=1}^{r} Y_i \sim Gamma(r, \lambda)$ (special case is called Erlang distribution).
- The special case of $Gamma(k/2, 1/2)$ is called the chi-square distribution with k degrees of freedom and denoted by χ^2_k. We will soon see its importance.

A random variable X is said to follow the normal distribution with parameters $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ (the *mean* and *variance* parameters, respectively), denoted $X \sim \mathrm{N}(\mu, \sigma^2)$, if

$$f_X(x;\mu,\sigma^2) = \frac{1}{\underset{1}{\sigma}\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{x-\overset{0}{\mu}}{\underset{1}{\sigma}}\right)^2\right\}, \quad x \in \mathbb{R}.$$

The mean, variance and moment generating function of $X \sim \mathrm{N}(\mu, \sigma^2)$ are given by

$$\mathbb{E}[X] = \mu, \qquad \mathrm{var}[X] = \sigma^2, \qquad M(t) = \exp\{t\mu + t^2\sigma^2/2\}.$$

In the special case $Z \sim \boxed{N(0,1)}$, we use the notation $\varphi(z) = f_Z(z)$ and $\Phi(z) = F_Z(z)$, and call these the *standard normal density* and *standard normal CDF*, respectively.

Lemma

Let $X \sim N(\mu, \sigma^2)$, $a \neq 0$. Then $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Condequently, if $X \sim N(\mu, \sigma^2)$, then

$$F_X(x) = \Phi\left(\frac{x \boxed{- \mu}}{\sigma}\right),$$

where Φ is the standard normal CDF, $\Phi(u) = \int_{-\infty}^{u} (2\pi)^{-1/2} \exp\{-z^2/2\} dz$.

Corollary

Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \sim N(\mu_i, \sigma_i^2)$, and let $S_n = \sum_{i=1}^{n} X_i$. Then,

$$S_n \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right).$$

Entropy

Can one probability model be “more disordered” than another?

Can one probability model be "more disordered" than another?

The entropy of a random variable X is defined as

$$H(X) = \underline{-\mathbb{E}\Big[\overbrace{\log f_X(X)}^{g(X)}\Big]} = \begin{cases} -\sum_{x\in\mathcal{X}} f_X(x)\log\left(f_X(x)\right), & \text{if } X \text{ is discrete,} \\ -\int_{-\infty}^{+\infty} f_X(x)\log\left(f_X(x)\right)dx, & \text{if } X \text{ is continuous.} \end{cases}$$

- A measure of the intrinsic disorder or unpredictability of a random system.
- Related to but not equivalent to variance.

When X is discrete:

- $H(X) \geq 0$
- $H(g(X)) \leq H(X)$ for any deterministic function g.

Can we use entropy to compare distributions?

KL Divergence (Relative Entropy)

Can we use entropy to compare distributions? $\|\cdot\|_2^2$

Let $p(x)$ and $q(x)$ be two probability density (frequency) functions on $\mathbb{R}$. We define the Kullback-Leibler divergence or relative entropy of q with respect to p as

$$KL(q\|p) := \int_{-\infty}^{+\infty} p(x)\log\left(\frac{p(x)}{q(x)}\right)dx. = \mathbb{E}_p\left[\log\left(\frac{p(x)}{q(x)}\right)\right]$$

- By Jensen's inequality, for $X \sim p(\cdot)$ we have

$$KL(q\|p) = \mathbb{E}[\log[q(X)/p(X)]] \geq -\log\left(\mathbb{E}\left[\frac{q(X)}{p(X)}\right]\right) = 0$$

  since q integrates to 1.
- $p = q \iff KL(q\|p) = 0$.
- $KL(q\|p) \neq KL(p\|q)$.
- Not a metric (lacks symmetry and violates triangle inequality).

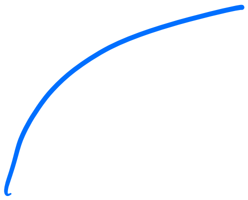
d

$d(x,z) \leq d(x,y) + d(y,z)$

$$KL(q \| p) = \mathbb{E}_p\left[\log\left(\frac{p}{q}\right)\right] \qquad \log(a^x) = x \cdot \log(a)$$

$$= \mathbb{E}_p\left[-\log\left(\frac{q}{p}\right)\right] \quad \left(\mathbb{E}_p\left[\log\left(\frac{p}{q}\right)^{-1}\right]\right)$$

$$=$$



Exponential Families

Consider the following variational problem:

Determine the probability distribution f supported on $\mathcal{X}$ with maximum entropy

$$f^* = \arg\max_f H(f) = -\int_{\mathcal{X}} f(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x}$$

subject to the linear constraints

$$\mathbb{E}_g[T_i] = \mathbb{E}_f[T_i(X)] = \int_{\mathcal{X}} T_i(\boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} \overset{\le}{=} \alpha_i, \quad i = 1, \ldots, k$$

Philosophy: How to choose a probability model for a given situation?

Maximum entropy approach:

- In any given situation, choose the distribution that gives *highest uncertainty* while satisfying situation–specific required constraints.

Proposition.

When a solution to the constrained optimisation problem exists, it is unique and has the form

$$f^*(\boldsymbol{x}) = Q(\lambda_1, \dots, \lambda_k) \exp\left\{\sum_{i=1}^{k} \lambda_i T_i(\boldsymbol{x})\right\} \qquad e^{\log Q}$$

Proof.

Let $g(\boldsymbol{x})$ be a density also satisfying the constraints. Then,

multiply & divide by f(x) inside log.

$$\begin{aligned} H(g) &\overset{\text{defn.}}{=} -\int_{\mathcal{X}} g(\boldsymbol{x}) \log g(\boldsymbol{x}) d\boldsymbol{x} = -\int_{\mathcal{X}} g(\boldsymbol{x}) \log\left[\frac{g(\boldsymbol{x})}{f(\boldsymbol{x})} f(\boldsymbol{x})\right] d\boldsymbol{x} \\ &= -\underbrace{KL(g \| f)}_{\geq 0} - \int_{\mathcal{X}} g(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x} \\ &\leq -\log Q \underbrace{\int_{\mathcal{X}} g(\boldsymbol{x}) dx}_{=1} - \int_{\mathcal{X}} g(\boldsymbol{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\boldsymbol{x})\right) d\boldsymbol{x} \end{aligned}$$

plugging in

$*\ \log(a \cdot b) = \log(a) + \log(b)$

$$1 \quad \int_{\mathcal{X}} g(x) \log \Bigg(\underbrace{\frac{g(x)}{f(x)}}_{a} \underbrace{f(x)}_{b} \Bigg) dx$$

$$\overset{*}{=} \int_{\mathcal{X}} g(x) \left[\log\left(\frac{g(x)}{f(x)}\right) + \log f(x) \right] dx$$

$$= \underbrace{\int_{\mathcal{X}} g(x) \log\left(\frac{g(x)}{f(x)}\right)}_{KL(f \| g)} + \underbrace{\int_{\mathcal{X}} g(x) \log f(x)\, dx}$$

$$= \int g(x) \log \Big(\underbrace{Q(\lambda_1, \ldots \lambda_k)}\, \underbrace{\exp(\textstyle\sum \lambda_i T_i)} \Big) dx$$

$$= \int g(x) \log \underline{Q}\, dx + \underline{\int g(x) \cancel{\log}\, \cancel{\exp}(\textstyle\sum \lambda_i T_i)\, dx}$$

$KL \geq 0 \Rightarrow g=f, KL=0 \Rightarrow H(g)=H(f)$

$H(g) \leq H(f)$

But g also satisfies the moment constraints, so the last term is

$$= \quad -\log Q - \int_{\mathcal{X}} f(\boldsymbol{x}) \left(\sum_{i=1}^{k} \lambda_i T_i(\boldsymbol{x}) \right) d\boldsymbol{x} = -\int_{\mathcal{X}} f(\boldsymbol{x}) \log f(\boldsymbol{x}) d\boldsymbol{x}$$

$\int f(x)dx$ $\quad g(x)$

defn & f^*

$$= \quad H(f)$$

Uniqueness of the solution follows from the fact that strict equality can only follow when $KL(g \| f) = 0$, which happens if and only if $g = f$. □

Exponential Family of Distributions

A probability distribution is said to be a member of a k-parameter exponential family, if its density (or frequency) admits the representation

independent of R.V.

$$f(y) = \exp\left\{\sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\phi_1, \ldots, \phi_k) + S(y)\right\}$$

log Q(- - -)

where:

1. $\phi = (\phi_1, \ldots, \phi_k)$ is a k-dimensional parameter in $\Phi \subseteq \mathbb{R}^k$;
2. $T_i : \mathcal{Y} \to \mathbb{R}$, $i = 1, \ldots, k$, $S : \mathcal{Y} \to \mathbb{R}$, and $\gamma : \mathbb{R}^k \to \mathbb{R}$, are real-valued;
3. The support $\mathcal{Y}$ of f does not depend on ϕ.

Very rich class of models (sometimes requiring fixing some parameters to satisfy last condition): Binomial, Negative Binomial, Poisson, Gamma, Gaussian, Pareto, Weibull, Laplace, logNormal, inverse Gaussian, inverse Gamma, Normal-Gamma, Beta, Multinomial...

$\hookrightarrow$ Basis for *Generalised Linear Models (GLM)*.

We will gradually see that such models have magnificent properties.

$\sigma^2 > 0$

- $\boldsymbol{\phi} = (\phi_1, ..., \phi_k)^\top$ is called the natural parameter
- But transforming parameter, we can write exponential family in other ways.
- "Natural" is from the mathematics point of view – usual parameter $\boldsymbol{\theta} = \eta^{-1}(\boldsymbol{\phi})$ often different.

Natural vs Usual Parametrization

$$\exp\left\{\sum_{i=1}^{k} \phi_i T_i(y) - \gamma(\boldsymbol{\phi}) + S(y)\right\} = \exp\left\{\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta}) T_i(y) - d(\boldsymbol{\theta}) + S(y)\right\}.$$

where $\eta : \mathbb{R}^k \to \mathbb{R}^k$ is a C^2 map such that

twice continuously differentiable

η''

$$\boldsymbol{\phi} = \eta(\boldsymbol{\theta})$$

and so $\gamma(\boldsymbol{\phi}) = \gamma(\eta(\boldsymbol{\theta})) = d(\boldsymbol{\theta})$, for $d = \gamma \circ \eta$.

- Natural parametrization: great for mathematical manipulation.
- Usual parametrization: more intuitive in context of applications.

Example (Binomial Exponential Family)

Let $Y \sim \text{Binom}(n, p)$. Observe that: $\log\left[\binom{n}{y} p^y (1-p)^{n-y}\right]$

$$f_Y(y) = \binom{n}{y} p^y (1-p)^{n-y} = \exp\left\{\log\left(\frac{p}{1-p}\right)\boxed{y} + n\log(1-p) + \underbrace{\log\binom{n}{y}}_{S(y)}\right\}.$$

$= T(y)$

Define (for fixed n)

$$\phi = \log\left(\frac{p}{1-p}\right), \quad T(y) = y,$$

$$S(y) = \log\binom{n}{y}, \quad \gamma(\phi) = n\log(1+e^{\phi}) = -n\log(1-p).$$

Keeping n fixed and allowing only p to vary, the support of f does not depend on ϕ and we get a 1-parameter exponential family. Note that:

$$p = \frac{e^{\phi}}{1+e^{\phi}} \qquad \& \qquad \phi = \underbrace{\log\left(\frac{p}{1-p}\right)}_{=\eta(p)}.$$

so the usual parameter is $p \in (0, 1)$, but the natural one is $\phi \in \mathbb{R}$. $\square$

$$\log\left[\binom{n}{y} p^{y}(1-p)^{n-y}\right] = \log\binom{n}{y} + y\log p + \underbrace{(n-y)}\log(1-p)$$

$$= \log\binom{n}{y} + y\log p + n\log(1-p) + y\log(1-p)^{-1}$$

$$\log\left(\frac{a}{b}\right) = \log a - \log b$$

$$= \log\binom{n}{y} + y\log\left(\frac{p}{1-p}\right) + n\log(1-p)$$

$$\log\left(\frac{p}{1-p}\right) = \phi$$

$$\frac{p}{1-p} = e^{\phi}$$

$$p = e^{\phi} - p e^{\phi}$$

$$p = \frac{e^{\phi}}{1+e^{\phi}}$$

Example (Gaussian Exponential Family)

Let $Y \sim N(\mu, \sigma^2)$. We can write

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

$f_\theta(y)$

$$= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}.$$

not dependent on y

Define

$\mathbb{E}[T_i(y)] = \mathbb{E}[y^2]$

$$\phi_1 = \frac{\mu}{\sigma^2}, \quad \phi_2 = -\frac{1}{2\sigma^2},$$

$$T_1(y) = y, \quad T_2(y) = y^2, \quad S(y) = 0, \quad \gamma(\phi_1, \phi_2) = -\frac{\phi_1^2}{4\phi_2} + \frac{1}{2}\log\left(-\frac{\pi}{\phi_2}\right),$$

and observe that the support of f is always $\mathbb{R}$. Thus $N(\mu, \sigma^2)$ is a two-parameter exponential family. □

Sampling Theory and Stochastic Convergence

data parameter(s)

1. Model phenomenon by distribution $F(y_1, \ldots, y_n; \theta)$ on $\boxed{\mathcal{Y}^n}$, some $n \geq 1$.

   $\theta = (\mu, \sigma^2)$
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, \ldots, Y_n)^\top \in \mathcal{Y}^n$ from this distribution. Call this a sample.
4. Use sample $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

$\int_{y} y\, g(y)\, dy = \mathbb{E}[y]$

1. Model phenomenon by distribution $F(y_1, ..., y_n; \theta)$ on $\mathcal{Y}^n$, some $n \geq 1$.
2. Distributional form is known but $\theta \in \Theta$ is unknown.
3. Observe realisation of $(Y_1, ..., Y_n)^\top \in \mathcal{Y}^n$ from this distribution. Call this a sample.
4. Use sample $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

$$\boxed{\text{Anything we do will be a function } T(Y_1, \ldots, Y_n) \text{ of the sample}}$$

Sampling theory aims to understand:

1. What information do different forms of functions $T : \mathcal{Y}^n \to \mathbb{R}^p$ carry on the parameter θ?
2. What is the probability distribution of $T(Y_1, ..., Y_n)$ and how does it relate to $F(y_1, ..., y_n; \theta)$?

These two questions are closely related.

$X = g(Y)$ $\mathbb{E}[Z|Y]$

Definition (Statistic)

A statistic is any function T whose domain is the sample space $\mathcal{Y}^n$ but does not depend on unknown parameters.

$\hookrightarrow$ Intuitively, any function that can be evaluated on the basis of the sample alone is a statistic.

$\hookrightarrow$ Any statistic is clearly itself a random variable with its own distribution.

Example

$T: \mathcal{Y}^n \to \mathbb{R}$

$T(\boldsymbol{Y}) = n^{-1}\sum_{i=1}^{n} Y_i$ is a statistic (since n, the sample size, is known).

Example

$T: \mathcal{Y}^n \to \mathbb{R}^n (\mathcal{Y}^n)$

$T(\boldsymbol{Y}) = (Y_{(1)}, \ldots, Y_{(n)})$ where $Y_{(1)} \leq Y_{(2)} \leq \ldots, Y_{(n)}$ are the order statistics of $\boldsymbol{Y}$. Since T depends only on the values of $\boldsymbol{Y}$, T is a statistic.

Example

Let $T(\boldsymbol{Y}) = c$, where c is a known constant. Then T is a statistic

Definition (Sampling Distribution)

Let $(Y_1, \ldots, Y_n)^\top \sim F(y_1, ..., y_n; \theta)$ and $T : \mathcal{Y}^n \to \mathbb{R}^q$ be a statistic,

$$T(Y_1, ..., Y_n) = (T_1(Y_1, \ldots, Y_n), \ldots, T_q(Y_1, \ldots, Y_n)).$$

The sampling distribution of T under $F(y_1, ..., y_n; \theta)$ is the distribution

$$F_T(t_1, \ldots, t_q) = \mathbb{P}[T_1(Y_1, \ldots, Y_n) \leq t_1, \ldots, T_q(Y_1, \ldots, Y_n) \leq t_q].$$

Comments:

- We will typically simply write T instead of the cumbersome $T(Y_1, ..., Y_n)$.
- Very often $T : \mathcal{Y}^n \to \mathbb{R}$ (i.e. $q = 1$), in which case the notation simplifies considerably:
$$F_T(t) = \mathbb{P}[T(Y_1, ..., Y_n) \leq t], \quad t \in \mathbb{R}.$$

Key observation:

$$\boxed{\text{The sampling distribution of } T \text{ depends on the unknown } \theta}$$

The extent and form of this dependence is essential for inference.

- Evident from previous examples: some statistics are more informative and others are less informative regarding the true value of θ
- Any $T(Y_1, ..., Y_n)$ that is not "1-1" carries less information about θ than the original sample $(Y_1, ..., Y_n)$ itself.
- Which are "good" and which are "bad" statistics?

Definition (Ancillary Statistic)

A statistic T is an *ancillary statistic* (for θ) if its distribution does not functionally depend θ

$\hookrightarrow$ So an ancillary statistic has the same distribution $\forall\ \theta \in \Theta$.

Example

Suppose that $Y_1, ..., Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where μ unknown but σ^2 known). Let $T(Y_1, ..., Y_n) = Y_1 - Y_2$; then T has a Normal distribution with mean 0 and variance $2\sigma^2$. Thus T is ancillary for the unknown parameter μ. If both μ and σ^2 were unknown, T would not be ancillary for $\boldsymbol{\theta} = (\mu, \sigma^2)$.

$T \sim \mathcal{N}(0, 2\sigma^2)$

- If T is ancillary for θ then T carries no information about θ
- In order to carry any useful information about θ, the sampling distribution F_T must depend explicitly on θ.
- Intuitively, the amount of information T carries on θ increases as the dependence of its sampling distribution F_T on θ increases

- If T is ancillary for θ then T carries no information about θ
- In order to carry any useful information about θ, the sampling distribution F_T must depend explicitly on θ.
- Intuitively, the amount of information T carries on θ increases as the dependence of its sampling distribution F_T on θ increases

Example

Let $Y_1, ..., Y_n \stackrel{iid}{\sim} \mathcal{U}[0, \theta]$, $S = \min(Y_1, \ldots, Y_n)$ and $T = \max(Y_1, \ldots, Y_n)$.

- $f_S(y; \theta) = \frac{n}{\theta}\left(1 - \frac{y}{\theta}\right)^{n-1}, \quad 0 \leq y \leq \theta$
- $f_T(y; \theta) = \frac{n}{\theta}\left(\frac{y}{\theta}\right)^{n-1}, \quad 0 \leq y \leq \theta$

$\left(\frac{a}{b}\right)^n \approx \left(\frac{b}{b}\right)^n + \varepsilon$

$\hookrightarrow$ Neither S nor T are ancillary for θ

$\hookrightarrow$ As $n \uparrow \infty$, f_S becomes concentrated around 0

$\hookrightarrow$ As $n \uparrow \infty$, f_T becomes concentrated around θ while

$\hookrightarrow$ Indicates that T provides more information about θ than does S.

$S = \min\{Y_1, \dots Y_n\}$

$$F_S(t) = \mathbb{P}(S \le t) = 1 - \mathbb{P}(S > t)$$

$$= \mathbb{P}(\min\{Y_1, \dots Y_n\} \le t)$$

$$= 1 - \mathbb{P}(\min\{Y_1, \dots Y_n\} > t)$$

$$= 1 - \mathbb{P}(Y_1 > t, \dots Y_n > t)$$

$$= 1 - \prod_{i=1}^{n} \mathbb{P}(Y_i > t)$$

$$\downarrow$$

Unif $[0, \theta]$