

Statistics for Data Science: Week 2

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

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



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:

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

Among all functions¹ of Y , $\mathbb{E}[X|Y]$ best approximates X in mean square.

¹measurable

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

- ① 'Tower property': $\mathbb{E}\left[\mathbb{E}[X|Y]\right] = \mathbb{E}[X]$
- 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: $X_1 \leq X_2 \implies \mathbb{E}[X_1|Y] \leq \mathbb{E}[X_2|Y]$

The **conditional variance** of X given Y is defined as

$$\text{var}[X|Y] = \mathbb{E}\left[(X - \mathbb{E}[X|Y])^2 \mid Y\right] = \mathbb{E}[X^2|Y] - (\mathbb{E}[X|Y])^2$$

The **law of total variance** states that

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

Proof:

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

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

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

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

If we write

$$\mu = \mathbb{E}[\mathbf{Y}] = (\mathbb{E}[Y_1], \dots, \mathbb{E}[Y_d])^\top$$

for the mean vector of \mathbf{Y} , then

$$\mathbb{E}[(\mathbf{Y} - \mu)(\mathbf{Y} - \mu)^\top] = \mathbb{E}[\mathbf{Y}\mathbf{Y}^\top] - \mu\mu^\top.$$

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

Let \mathbf{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 \mathbf{A} is a $p \times d$ deterministic matrix, the mean vector and covariance matrix of \mathbf{AY} are $\mathbf{A}\boldsymbol{\mu}$ and $\mathbf{A}\boldsymbol{\Omega}\mathbf{A}^\top$, respectively.
- If $\boldsymbol{\beta} \in \mathbb{R}^d$ is a deterministic vector, the variance of $\boldsymbol{\beta}^\top \mathbf{Y}$ is $\boldsymbol{\beta}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.
- If $\boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$ are deterministic vectors, the covariance of $\boldsymbol{\beta}^\top \mathbf{Y}$ with $\boldsymbol{\gamma}^\top \mathbf{Y}$ is $\boldsymbol{\gamma}^\top \boldsymbol{\Omega} \boldsymbol{\beta}$.

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

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

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

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

For any convex² function $\varphi : \mathbb{R} \rightarrow \mathbb{R}$, if $\mathbb{E}|\varphi(X)| + \mathbb{E}|X| < \infty$, then one has

$$\varphi\left(\mathbb{E}[X]\right) \leq \mathbb{E}[\varphi(X)] \quad [\text{Jensen}]$$

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

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

²Recall 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 $\lambda \in [0, 1]$.

Moment Generating Functions

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

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

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

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 \mathbf{X} in \mathbb{R}^d , the MGF is

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

$$M_{\mathbf{X}}(\mathbf{u}) = \mathbb{E}[e^{\mathbf{u}^\top \mathbf{X}}], \quad \mathbf{u} \in \mathbb{R}^d.$$

and has analogous properties.

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

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

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

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 $\mathcal{X} = \{0, 1, 2, \dots, n\}$,

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

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

- If $X = \sum_{i=1}^n Y_i$ where $Y_i \stackrel{iid}{\sim} \text{Bern}(p)$, 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 $X \sim \text{Geom}(p)$, if

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

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

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

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

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

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

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

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

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

- If $X = \sum_{i=1}^r Y_i$ where $Y_i \stackrel{iid}{\sim} \text{Geom}(p)$, 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 $f(x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$.

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

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

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

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

① the sample space is $\{0, 1, \dots, n\}^k$, and

② $f(x_1, \dots, x_k; n, \{p_i\}_{i=1}^k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \dots p_k^{x_k} \mathbf{1} \left\{ \sum_{i=1}^k x_i = n \right\}.$

The mean, variance covariance and moment generating function are

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

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

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

Lemma (Poisson and Multinomial)

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

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

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

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

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

$$\mathbb{E}[X] = (\theta_1 + \theta_2)/2, \quad \text{var}[X] = (\theta_2 - \theta_1)^2/12, \quad M(t) = \frac{e^{t\theta_2} - e^{t\theta_1}}{t(\theta_2 - \theta_1)}, \quad t \neq 0$$
$$M(0) = 1.$$

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

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

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 $X \sim \text{Exp}(\lambda)$. 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

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

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

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

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

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

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

- If $Y_1, \dots, Y_r \stackrel{iid}{\sim} \text{Exp}(\lambda)$, then $Y = \sum_{i=1}^r Y_i \sim \text{Gamma}(r, \lambda)$ (special case is called Erlang distribution).
- The special case of $\text{Gamma}(k/2, 1/2)$ is called the **chi-square distribution with k degrees of freedom** and denoted by χ_k^2 . 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 $X \sim N(\mu, \sigma^2)$, if

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

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

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

In the special case $Z \sim N(0, 1)$, 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 $aX + b \sim N(a\mu + b, a^2\sigma^2)$. Consequently, if $X \sim N(\mu, \sigma^2)$, then

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

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, \dots, 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?

The **entropy** of a random variable X is defined as

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

- 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?

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

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

- 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).

Exponential Families

Consider the following variational problem:

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

$$H(f) = - \int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x}$$

subject to the linear constraints

$$\int_{\mathcal{X}} T_i(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = \alpha_i, \quad i = 1, \dots, k$$

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

$$f(\mathbf{x}) = Q(\phi_1, \dots, \phi_k) \exp \left\{ \sum_{i=1}^k \phi_i T_i(\mathbf{x}) \right\}$$

Proof.

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

$$\begin{aligned} H(g) &= - \int_{\mathcal{X}} g(\mathbf{x}) \log g(\mathbf{x}) d\mathbf{x} = - \int_{\mathcal{X}} g(\mathbf{x}) \log \left[\frac{g(\mathbf{x})}{f(\mathbf{x})} f(\mathbf{x}) \right] d\mathbf{x} \\ &= - \underbrace{KL(g \parallel f)}_{\geq 0} - \int_{\mathcal{X}} g(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &\leq - \underbrace{\log Q \int_{\mathcal{X}} g(\mathbf{x}) d\mathbf{x}}_{=1} - \int_{\mathcal{X}} g(\mathbf{x}) \left(\sum_{i=1}^k \phi_i T_i(\mathbf{x}) \right) d\mathbf{x} \end{aligned}$$

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

$$\begin{aligned} &= -\log Q - \int_{\mathcal{X}} f(\mathbf{x}) \left(\sum_{i=1}^k \phi_i T_i(\mathbf{x}) \right) d\mathbf{x} = - \int_{\mathcal{X}} f(\mathbf{x}) \log f(\mathbf{x}) d\mathbf{x} \\ &= H(f) \end{aligned}$$

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

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

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

where:

- ① $\phi = (\phi_1, \dots, \phi_k)$ is a k -dimensional parameter in $\Phi \subseteq \mathbb{R}^k$;
- ② $T_i : \mathcal{Y} \rightarrow \mathbb{R}$, $i = 1, \dots, k$, $S : \mathcal{Y} \rightarrow \mathbb{R}$, and $\gamma : \mathbb{R}^k \rightarrow \mathbb{R}$, 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...

→ Basis for *Generalised Linear Models (GLM)*.

We will gradually see that such models have magnificent properties.

- $\phi = (\phi_1, \dots, \phi_k)^\top$ 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**
 $\theta = \eta^{-1}(\phi)$ often different.

Natural vs Usual Parametrization

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

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

$$\phi = \eta(\theta)$$

and so $\gamma(\phi) = \gamma(\eta(\theta)) = d(\theta)$, 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:

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

Define

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

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

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:

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

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

Example (Gaussian Exponential Family)

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

$$\begin{aligned}f(y; \mu, \sigma^2) &= \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right\} \\&= \exp\left\{-\frac{1}{2\sigma^2}y^2 + \frac{\mu}{\sigma^2}y - \frac{1}{2}\log(2\pi\sigma^2) - \frac{\mu^2}{2\sigma^2}\right\}.\end{aligned}$$

Define

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

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

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

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

Anything we do will be a function $T(Y_1, \dots, Y_n)$ of the sample

Sampling theory aims to understand:

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

These two questions are closely related.

Definition (Statistic)

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

- ↪ Intuitively, any function that can be evaluated on the basis of the sample alone is a statistic.
- ↪ Any statistic is clearly itself a random variable with its own distribution.

Example

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

Example

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

Example

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

Definition (Sampling Distribution)

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

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

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

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

Comments:

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

$$F_T(t) = \mathbb{P}[T(Y_1, \dots, Y_n) \leq t], \quad t \in \mathbb{R}.$$

Key observation:

The sampling distribution of T depends on the unknown θ

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 $T(Y_1, \dots, Y_n)$ that is not “1-1” carries less information about θ than the original sample (Y_1, \dots, 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

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

Example

Suppose that $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ (where μ unknown but σ^2 known). Let $T(Y_1, \dots, Y_n) = Y_1 - Y_2$; 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 $\theta = (\mu, \sigma^2)$.

- 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

Example

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

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

- ↪ Neither S nor T are ancillary for θ
- ↪ As $n \uparrow \infty$, f_S becomes concentrated around 0
- ↪ As $n \uparrow \infty$, f_T becomes concentrated around θ while
- ↪ Indicates that T provides more information about θ than does S .