Regularity conditions

- These conditions are sufficient (but not necessary) to prove theorems giving the limiting distributions for $\widehat{\theta}$ and W as the sample size (or more generally some measure of the information in the data) goes to infinity.
- Why they are needed:
  - (C1) ensures that $\widehat{\theta}$ can be 'on all sides' of θ^0 in the limit — if it fails, then any limiting distribution cannot be normal;
  - (C2) is essential for consistency, otherwise $\widehat{\theta}$ might not converge to a unique limit;
  - (C3) is needed to bound terms of a Taylor series — can be replaced by other conditions; and
  - (C4) ensures that $\widehat{\theta}$ is consistent for θ^0 and that the asymptotic variance of $\widehat{\theta}$ is the inverse Fisher information $\imath(\theta^0)^{-1}$.
- In some of the models arising later, (C4) may fail (or be close to failing), because the support of the data depends on a parameter.

MGFs and KGFs

- The **moment-generating function (MGF)** and **cumulant-generating function (KGF)** of a scalar random variable X are

$$M_X(t) = \mathrm{E}\left(e^{tX}\right), \quad K_X(t) = \log M_X(t), \quad t \in \mathcal{N} = \{t : M_X(t) < \infty\}.$$

- $\mathcal{N}$ is non-empty, because $M_X(0) = 1$, but the MGF and KGF are non-trivial only if $\mathcal{N}$ contains an open neighbourhood of the origin, since then

$$M_X(t) = \mathrm{E}\left(\sum_{r=0}^{\infty} \frac{t^r X^r}{r!}\right) = \sum_{r=0}^{\infty} \frac{t^r}{r!}\mathrm{E}(X^r), \quad K_X(t) = \sum_{r=1}^{\infty} \frac{t^r}{r!}\kappa_r,$$

  and one can obtain the **moments** $\mathrm{E}(X^r)$ and **cumulants** κ_r by differentiation.
- If $X = (X_1, \ldots, X_n)$ then we set $t = (t_1, \ldots, t_n)^{\mathrm{T}}$ and define $M_X(t) = \mathrm{E}\left(e^{t^{\mathrm{T}}X}\right)$ and $K_X(t)$ as above. Then the mean and covariance matrix of X are $\mathrm{E}(X) = \nabla_t K_X(t)|_{t=0}$ and $\mathrm{cov}(X) = \nabla_t^2 K_X(t)|_{t=0}$.
- There is a 1–1 mapping between distributions and MGFs/KGFs (if the latter are non-trivial).
- If $X = (X_1, X_2)$, $t = (t_1, t_2)$ then

$$M_X(t) = \mathrm{E}\left(e^{t^{\mathrm{T}}X}\right) = E\left(e^{t_1^{\mathrm{T}}X_1}\right) E\left(e^{t_2^{\mathrm{T}}X_2}\right) = M_{X_1}(t)M_{X_2}(t), \forall t \quad \Longleftrightarrow \quad X_1 \perp\!\!\!\perp X_2.$$

Note: Moments and cumulants

- ☐ We consider scalar X, as the calculations for vector X are analogous.
- ☐ First note that $M_X(t) = 1$ when $t = 0$, since $\mathrm{E}(e^{tX}) = \mathrm{E}(1) = 1$; thus $0 \in \mathcal{N}$ for any X.
- ☐ If $\mathcal{N}$ contains an open set $(-a, a)$ for some $a > 0$, and $\mu_r = \mathrm{E}(X^r)$ denotes the rth moment of X, then if $|t| < a$,

$$K_X(t) = \sum_{r=1}^{\infty} \frac{t^r \kappa_r}{r!} = \log M_X(t) = \log\left(\sum_{r=0}^{\infty} \frac{t^r \mu_r}{r!}\right) = \log(1+b) = b - b^2/2 + b^3/3 + \cdots,$$

where $b = t\mu_1 + t^2\mu_2/2! + t^3\mu_3/3! + \cdots$. If we expand and compare coefficients of $t, t^2, t^3, \ldots$ in the two expansions we get

$$\kappa_1 = \mu_1, \quad \kappa_2 = \mu_2 - \mu_1^2, \quad \kappa_3 = \mu_3 - 3\mu_2\mu_1 + 2\mu_1^3, \quad \kappa_4 = \mu_4 - 4\mu_3\mu_1 + 6\mu_2\mu_1^2 - 3\mu_1^4, \quad \ldots,$$

so $\kappa_1 = \mathrm{E}(X)$, $\kappa_2 = \mathrm{var}(X)$, $\kappa_3 = \mathrm{E}\{(X - \mu_1)^3\}$, ...

Exponential family models

- ☐ If $\theta \in \Theta \subset \mathbb{R}^d$, where $\dim \Theta = d$, and there exists a $d \times 1$ function $s = s(y)$ of data y and a **parametrisation** (i.e., a 1–1 function) $\varphi \equiv \varphi(\theta)$ such that

$$f(y;\theta) = m(y) \exp\{s^{\mathrm{T}}\varphi - k(\varphi)\} = m(y) \exp\left[s^{\mathrm{T}}\varphi(\theta) - k\{\varphi(\theta)\}\right], \quad \theta \in \Theta, y \in \mathcal{Y},$$

then this is an (d, d) **exponential family** of distributions, with

  - **canonical statistic** $S = s(Y)$,
  - **canonical parameter** φ,
  - **cumulant generator** k, which is convex on $\mathcal{N} = \{\varphi : k(\varphi) < \infty\}$, and
  - **mean parameter** $\mu \equiv \mu(\varphi) = \mathrm{E}(S;\varphi) = \nabla_\varphi k(\varphi)$.

- ☐ We suppose that there is no vector $a \neq 0$ such that $a^{\mathrm{T}}S$ is constant, and call the model a **minimal representation** if there is no vector $a \neq 0$ such that $a^{\mathrm{T}}\varphi$ is constant.
- ☐ The cumulant-generating function for S is

$$K_S(t) = \log M_S(t) = k(\varphi + t) - k(\varphi), \quad t \in \mathcal{N}' \subset \mathbb{R}^d,$$

where $0 \in \mathcal{N}'$. One can check that

$$\mathrm{E}(S) = \nabla_\varphi k(\varphi), \quad \mathrm{var}(S) = \nabla_\varphi^2 k(\varphi).$$

Note: Cumulant-generating functions

☐ The MGF for the canonical statistic S of an exponential family is

$$M_S(t) = \mathrm{E}\left\{\exp(t^{\mathrm{T}}S)\right\} = \int m(y)\exp\left\{s^{\mathrm{T}}t + s^{\mathrm{T}}\varphi - k(\varphi)\right\}\,\mathrm{d}y,$$

and since this must equal unity when $t = 0$ we see that

$$\int m(y)\exp\left\{s^{\mathrm{T}}\varphi\right\}\,\mathrm{d}y = \exp\{k(\varphi)\},$$

and therefore that if $t + \varphi \in \mathcal{N}$, then

$$M_S(t) = \int m(y)\exp\left\{s^{\mathrm{T}}(t+\varphi) - k(\varphi)\right\}\,\mathrm{d}y = \exp\{k(\varphi+t) - k(\varphi)\},$$

which yields $K_S(t) = k(\varphi + t) - k(\varphi)$.

☐ Now $M_S(0) = 1$, $K_S(0) = 0$, $\partial K_S(t)/\partial t = \nabla_\varphi k(\varphi + t)$ and $\partial^2 K_S(t)/\partial t\partial t^{\mathrm{T}} = \nabla^2_\varphi k(\varphi + t)$, so

$$\mathrm{E}(S) = \partial M_S(t)/\partial t|_{t=0} = \partial e^{K_S(t)}/\partial t\Big|_{t=0} = \partial K_S(t)/\partial t\, e^{K_S(t)}\Big|_{t=0} = \nabla_\varphi k(\varphi).$$

A similar calculation for the variance gives

$$\mathrm{E}(SS^{\mathrm{T}}) = \partial^2 M_S(t)/\partial t\partial t^{\mathrm{T}}\big|_{t=0} = \nabla^2_\varphi k(\varphi) + \nabla_\varphi k(\varphi)\nabla_\varphi k(\varphi)^{\mathrm{T}},$$

and thus

$$\mathrm{var}(S) = \mathrm{E}(SS^{\mathrm{T}}) - \mathrm{E}(S)\mathrm{E}(S)^{\mathrm{T}} = \nabla^2_\varphi k(\varphi) + \nabla_\varphi k(\varphi)\nabla_\varphi k(\varphi)^{\mathrm{T}} - \nabla_\varphi k(\varphi)\nabla_\varphi k(\varphi)^{\mathrm{T}} = \nabla^2_\varphi k(\varphi).$$

Examples

Example 1 (Poisson distribution) *If $Y \sim \mathrm{Poiss}(\lambda)$, find its MGF and KGF and give its mean and variance. Show also that a sum of independent Poisson variables has a Poisson distribution.*

Example 2 (Poisson sample) *If $Y_1, \ldots, Y_n \overset{\mathrm{iid}}{\sim} \mathrm{Poiss}(\lambda)$, find the corresponding exponential family.*

Example 3 (Log-linear model) *If $Y_1, \ldots, Y_n \overset{\mathrm{ind}}{\sim} \mathrm{Poiss}(\lambda_j)$, where $\lambda_j = \exp(\beta_0 + \beta_1 x_j)$ and $x_1, \ldots, x_n$v are known constants, find the corresponding exponential family.*

☐ **NB:** avoid confusion — exponential family $\neq$ exponential distribution! The exponential distribution is just one example of an exponential family.

Note to Example 1

- The probability mass function of X is $\mathrm{P}(X = x) = \lambda^x e^{-\lambda}/x!$, $x \in \{0, 1, \ldots\}$, $\lambda > 0$, so

$$M_X(t) = \mathrm{E}(e^{tX}) = \sum_{x=0}^{\infty} e^{xt} \frac{\lambda^x}{x!} e^{-\lambda} = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda e^t)^x}{x!} = \exp\{\lambda(e^t - 1)\}, \quad t \in \mathbb{R},$$

  so $K_X(t) = \lambda(e^t - 1)$, $\mathcal{N} = \mathbb{R}$. Thus $\kappa_r = \lambda$ for all $r = 1, 2, \ldots$, and in particular $\mathrm{E}(X) = \mathrm{var}(X) = \lambda$.

- Suppose that $X_1, \ldots, X_n$ are independent Poisson variables with respective means $\lambda_1, \ldots, \lambda_n$. Then for any $t \in \mathbb{R}$ their sum $S = X_1 + \cdots + X_n$ has MGF

$$\mathrm{E}(e^{tS}) = \mathrm{E}\{e^{t(X_1 + \cdots + X_n)}\} = \mathrm{E}\left\{\prod_{j=1}^{n} e^{tX_j}\right\} = \prod_{j=1}^{n} \mathrm{E}(e^{tX_j}) = \exp\left\{(\lambda_1 + \cdots + \lambda_n)(e^t - 1)\right\},$$

  where the third equality applies due to independence of the X_j. The bijection between MGFs and distributions implies that $S \sim \mathrm{Poiss}(\sum \lambda_j)$.

Note to Example 2

Independent Poisson $Y_1, \ldots, Y_n$ have joint probability mass function

$$f(y_1, \ldots, y_n; \lambda) = \prod_{j=1}^{n} f(y_j; \lambda) = \prod_{j=1}^{n} \frac{\lambda^{y_j}}{y_j!} e^{-\lambda} = m(y) \exp(s \log \lambda - n\lambda),$$

defined on $\mathcal{Y} = \{0, 1, 2, \ldots\}^n$, with $m(y) = (\prod y_j)^{-1}$. This is clearly a minimal representation of a $(1, 1)$ exponential family with

- canonical statistic $s = s(y) = \sum y_j$,
- canonical parameter $\log \lambda = \varphi \in \mathcal{N} = \mathbb{R}$,
- cumulant generator $k(\varphi) = n\lambda = ne^{\varphi}$ and
- mean parameter $\mu = \nabla_{\varphi} k(\varphi) = ne^{\varphi} = n\lambda = \mathrm{E}(S)$.

Two standard parametrizations use the real parameter φ or the mean $\mu = ne^{\varphi} \in \mathbb{R}_+$.

Note to Example 3

Here $Y_1, \ldots, Y_n$ have joint probability mass function

$$\prod_{j=1}^{n} f(y_j; \lambda_j) = \prod_{j=1}^{n} \frac{\lambda_j^{y_j}}{y_j!} e^{-\lambda_j} = \prod_{j=1}^{n} \frac{e^{\beta_0 y_j + \beta_1 x_j y_j}}{y_j!} e^{-e^{\beta_0 + x_j \beta_1}} = m(y) \exp\left\{s^{\mathrm{T}} \beta - k(\beta)\right\},$$

defined on $(y_1, \ldots, y_n) \in \mathcal{Y}$ and again with $m(y) = (\prod y_j)^{-1}$. This is a $(2, 2)$ exponential family with

- ☐ canonical statistic $s = s(y) = (\sum_j y_j, \sum_j x_j y_j)$,
- ☐ canonical parameter $\beta = (\beta_0, \beta_1) = \varphi \in \mathcal{N} = \mathbb{R}^2$,
- ☐ cumulant generator $k(\varphi) = \sum_j \exp(\beta_0 + x_j \beta_1)$ and
- ☐ mean parameter $\mu = \nabla_\varphi k(\varphi) = (\sum_j \exp(\beta_0 + x_j \beta_1), \sum_j x_j \exp(\beta_0 + x_j \beta_1)) = \mathrm{E}(S)$.

Provided that not all the x_j are equal, this is a minimal representation.

Exponential family models II

- ☐ Exponential families are **closed under sampling**: the joint density of independent observations $Y_1, \ldots, Y_n$ from an exponential family with the same $s(Y_j)^{\mathrm{T}} \varphi = S_j^{\mathrm{T}} \varphi$ is

$$\prod_{j=1}^{n} f(y_j; \theta) = \prod_{j=1}^{n} m(y_j) \exp\left\{s_j^{\mathrm{T}} \varphi - k_j(\varphi)\right\} = \prod_{j=1}^{n} m(y_j) \exp\left\{\left(\sum_{j=1}^{n} s_j\right)^{\mathrm{T}} \varphi - \sum_{j=1}^{n} k_j(\varphi)\right\},$$

  so with $k_S(\varphi) = \sum_j k_j(\varphi)$, the density of $S = \sum_j S_j = \sum_j s(Y_j)$ is

$$f(s; \theta) = m^*(s) e^{s^{\mathrm{T}} \varphi - k_S(\varphi)}, \quad \text{with} \quad m^*(s) = \int_{\{y: \sum_j s(y_j) = s\}} \prod_{j=1}^{n} m(y_j)\, \mathrm{d}y.$$

  This is an exponential family, with canonical statistic S, canonical parameter φ and cumulant generator $k_S(\varphi)$.

- ☐ The corresponding log likelihood function is

$$\ell(\varphi) = \log f(s; \theta) \equiv s^{\mathrm{T}} \varphi - k_S(\varphi),$$

  where $\equiv$ means that additive constants have been dropped, so

$$\nabla_\varphi \ell(\varphi) = s - \nabla_\varphi k_S(\varphi) = s - \mathrm{E}(S; \varphi), \quad \nabla_\varphi^2 \ell(\varphi) = -\nabla_\varphi^2 k_S(\varphi) = -\imath(\varphi),$$

  so the MLE $\widehat{\varphi}$ sets $\mathrm{E}(S; \varphi)$ equal to the observed s, and $\imath(\varphi) = \jmath(\varphi) = \mathrm{cov}(S; \varphi)$.