

Statistics for Data Science: Week 3

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

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



ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

To understand the information carried by statistics on θ , we need to understand how they partition the sample space \mathcal{Y}^n .

- $\mathbf{Y} = (Y_1, \dots, Y_n) \stackrel{iid}{\sim} F_\theta$ and $T(\mathbf{Y})$ a statistic.
- The *level sets* or *contours* of T are the sets

$$A_t = \{\mathbf{y} \in \mathcal{Y}^n : T(\mathbf{y}) = t\}.$$

(all potential samples that could have given us the value t for T)

↪ Clearly, T is constant when restricted to a level set.

- Any realization of \mathbf{Y} that falls in a given level set is equivalent as far as T is concerned, as T reduces all these values to the same output.
- Any inference drawn through T will be the same within a given level set.
- So let's look at the distribution of \mathbf{Y} conditional on a given level set A_t of T , $F_{\mathbf{Y}|T=t}(\mathbf{y}) \dots$

- Suppose $F_{\mathbf{Y}|T=t}$ changes depending on θ : we are losing information.
- Suppose $F_{\mathbf{Y}|T=t}$ is functionally independent of θ
 - \Rightarrow Then \mathbf{Y} contains no information about θ on the set A_t
 - \Rightarrow In other words, \mathbf{Y} is ancillary for θ on A_t
- If this is true for each $t \in \text{Range}(T)$ then $T(\mathbf{Y})$ contains the same information about θ as \mathbf{Y} itself does.
 - \hookrightarrow It does not matter whether we observe $\mathbf{Y} = (Y_1, \dots, Y_n)$ or just $T(\mathbf{Y})$.
 - \hookrightarrow Knowing the exact value \mathbf{Y} in addition to knowing $T(\mathbf{Y})$ does not give us any additional information - \mathbf{Y} is irrelevant if we already know $T(\mathbf{Y})$.

Definition (Sufficient Statistic)

A statistic $T = T(\mathbf{Y})$ is said to be *sufficient* for the parameter θ if the conditional probability distribution of the sample given the statistic

$$F_{\mathbf{Y}|T(\mathbf{Y})=t}(y_1, \dots, Y_n) = \mathbb{P}[Y_1 \leq y_1, \dots, Y_n \leq y_n | T(Y_1, \dots, Y_n) = t]$$

does *not* depend on θ .

Example (Coin Tossing)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$, and $T(\mathbf{Y}) = \sum_{i=1}^n Y_i$. For $\mathbf{y} \in \{0, 1\}^n$,

$$\begin{aligned}\mathbb{P}[\mathbf{Y} = \mathbf{y} | T = t] &= \frac{\mathbb{P}[\mathbf{Y} = \mathbf{y}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[\mathbf{Y} = \mathbf{y}]}{\mathbb{P}[T = t]} \mathbf{1}_{\{\sum_{i=1}^n y_i = t\}} \\ &= \frac{\theta^{\sum_{i=1}^n y_i} (1 - \theta)^{n - \sum_{i=1}^n y_i}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} \mathbf{1}_{\{\sum_{i=1}^n y_i = t\}} \\ &= \frac{\theta^t (1 - \theta)^{n-t}}{\binom{n}{t} \theta^t (1 - \theta)^{n-t}} \mathbf{1}_{\{\sum_{i=1}^n y_i = t\}} \\ &= \binom{n}{t}^{-1} \mathbf{1}_{\{\sum_{i=1}^n y_i = t\}}.\end{aligned}$$

- T is sufficient for $\theta \rightarrow$ Given $\#$ of tosses that came heads, knowing *which* tosses came heads is irrelevant in deciding the probability of heads:

0 0 1 1 1 0 1 VS 1 0 0 0 1 1 1 VS 1 0 1 0 1 0 1

- Definition hard to verify (especially for continuous variables)
- Definition does not allow easy identification of sufficient statistics

Theorem (Fisher-Neyman Factorization Theorem)

Suppose that $\mathbf{Y} = (Y_1, \dots, Y_n)$ has a joint density or frequency function $f(\mathbf{y}; \theta)$, $\theta \in \Theta$. A statistic $T = T(\mathbf{Y})$ is sufficient for θ if and only if

$$f(\mathbf{y}; \theta) = g(T(\mathbf{y}), \theta)h(\mathbf{y}).$$

Example

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{U}[0, \theta]$ with pdf $f(y; \theta) = \mathbf{1}\{y \in [0, \theta]\}/\theta$. Then,

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\theta^n} \mathbf{1}\{\mathbf{y} \in [0, \theta]^n\} = \frac{\mathbf{1}\{\max[y_1, \dots, y_n] \leq \theta\} \mathbf{1}\{\min[y_1, \dots, y_n] \geq 0\}}{\theta^n}$$

Therefore $T(\mathbf{Y}) = Y_{(n)} = \max[Y_1, \dots, Y_n]$ is sufficient for θ .

Proof of Neyman-Fisher Theorem - Discrete Real Statistic.

Suppose first that T is sufficient. Then

$$\begin{aligned}f(y; \theta) &= \mathbb{P}_\theta[\mathbf{Y} = \mathbf{y}] = \sum_t \mathbb{P}_\theta[\mathbf{Y} = \mathbf{y}, T = t] \\&= \mathbb{P}_\theta[\mathbf{Y} = \mathbf{y}, T = T(\mathbf{y})] = \mathbb{P}_\theta[T = T(\mathbf{y})]\mathbb{P}[\mathbf{Y} = \mathbf{y} | T = T(\mathbf{y})]\end{aligned}$$

Since T is sufficient, $\mathbb{P}[\mathbf{Y} = \mathbf{y} | T = T(\mathbf{y})]$ is independent of θ and so $f(y; \theta) = g(T(\mathbf{y}); \theta)h(\mathbf{y})$. Now suppose that $f(y; \theta) = g(T(\mathbf{y}); \theta)h(\mathbf{y})$. Then if $T(\mathbf{y}) = t$,

$$\begin{aligned}\mathbb{P}[\mathbf{Y} = \mathbf{y} | T = t] &= \frac{\mathbb{P}[\mathbf{Y} = \mathbf{y}, T = t]}{\mathbb{P}[T = t]} = \frac{\mathbb{P}[\mathbf{Y} = \mathbf{y}]}{\mathbb{P}[T = t]} \mathbf{1}\{T(\mathbf{y}) = t\} \\&= \frac{g(T(\mathbf{y}); \theta)h(\mathbf{y})\mathbf{1}\{T(\mathbf{y}) = t\}}{\sum_{\mathbf{z}: T(\mathbf{z})=t} g(T(\mathbf{z}); \theta)h(\mathbf{z})} = \frac{h(\mathbf{y})\mathbf{1}\{T(\mathbf{y}) = t\}}{\sum_{T(\mathbf{z})=t} h(\mathbf{z})}.\end{aligned}$$

which does not depend on θ . □

Example (Sufficient statistics for i.i.d. normal samples)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$. Recall that we can write

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

and so

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^n y_i^2 + \frac{\mu}{\sigma^2}\sum_{i=1}^n y_i - \frac{n}{2}\log(2\pi\sigma^2) - \frac{n\mu^2}{2\sigma^2}\right\}.$$

Consequently, Fisher-Neyman factorisation implies that the statistic

$$S(\mathbf{Y}) = (S_1(\mathbf{Y}), S_2(\mathbf{Y}))^\top = \left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i^2\right)^\top = (\bar{Y}, \sum_{i=1}^n Y_i^2)^\top$$

is sufficient for the parameter (μ, σ^2) and so is the statistic

$$T(\mathbf{Y}) = (T_1(\mathbf{Y}), T_2(\mathbf{Y}))^\top = \left(n^{-1}\sum_{i=1}^n Y_i, n^{-1}\sum_{i=1}^n (Y_i - \bar{Y})^2\right)^\top$$

since T and S are 1-1 functions of each other.

Example (Sufficient statistics for k -parameter exponential families)

More generally, consider a k -parameter exponential family, with density

$$f(y) = \exp \left\{ \sum_{j=1}^k \phi_j T_j(y) - \gamma(\phi_1, \dots, \phi_k) + S(y) \right\}, \quad y \in \mathcal{Y}.$$

Then an i.i.d. sample $(Y_1, \dots, Y_n)^\top$ has joint distribution

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \exp \left\{ \sum_{j=1}^k \phi_j \tau_j(y_1, \dots, y_n) - n\gamma(\phi_1, \dots, \phi_k) + \sum_{i=1}^n S(y_i) \right\}$$

where

$$\tau_j(y_1, \dots, y_n) = \sum_{i=1}^n T_j(y_i).$$

So the statistic

$$\tau(Y_1, \dots, Y_n) = (\tau_1(Y_1, \dots, Y_n), \dots, \tau_k(Y_1, \dots, Y_n))^\top$$

is sufficient for (ϕ_1, \dots, ϕ_k) by Fisher-Neyman factorisation.

- We have seen that sufficient statistic compresses data without information loss on parameter of interest.
- How much info can we throw away? Is there a “necessary” statistic?

Definition (Minimally Sufficient Statistic)

A statistic $T = T(\mathbf{Y})$ is said to be *minimally sufficient* for the parameter θ if it is sufficient for θ and for any other sufficient statistic $S = S(\mathbf{Y})$ there exists a function $g(\cdot)$ with

$$T(\mathbf{Y}) = g(S(\mathbf{Y})).$$

Lemma

If T and S are minimally sufficient statistics for a parameter θ , then there exists injective functions g and h such that $S = g(T)$ and $T = h(S)$.

Theorem

Let $\mathbf{Y} = (Y_1, \dots, Y_n)$ have joint density or frequency function $f(\mathbf{y}; \theta)$ and $T = T(\mathbf{Y})$ be a statistic. If $f(\mathbf{y}; \theta)/f(\mathbf{z}; \theta) \perp \theta \iff T(\mathbf{y}) = T(\mathbf{z})$. Then T is minimally sufficient for θ .

Proof.(*)

Assume for simplicity that $f(\mathbf{y}; \theta) > 0$ for all $\mathbf{y} \in \mathbb{R}^n$ and $\theta \in \Theta$. Let $\mathcal{T} = \{T(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^n\}$ be the image of \mathbb{R}^n under T and let A_t be the level sets of T . For each t , choose a representative element $\mathbf{w}_t \in A_t$. Notice that for any \mathbf{y} , $\mathbf{w}_{T(\mathbf{y})}$ is in the same level set as \mathbf{y} , so that

$$f(\mathbf{y}; \theta)/f(\mathbf{w}_{T(\mathbf{y})}; \theta)$$

does not depend on θ by assumption. Let $g(t, \theta) := f(\mathbf{w}_t; \theta)$ and notice

$$f(\mathbf{y}; \theta) = \frac{f(\mathbf{w}_{T(\mathbf{y})}; \theta)f(\mathbf{y}; \theta)}{f(\mathbf{w}_{T(\mathbf{y})}; \theta)} = g(T(\mathbf{y}), \theta)h(\mathbf{y})$$

and sufficiency follows from the Fisher-Neyman factorization theorem.

[minimality part] Suppose that T' is another sufficient statistic. By the factorization thm: $\exists g', h' : f(\mathbf{y}; \theta) = g'(T'(\mathbf{y}); \theta)h'(\mathbf{y})$. Let \mathbf{y}, \mathbf{z} be such that $T'(\mathbf{y}) = T'(\mathbf{z})$. Then

$$\frac{f(\mathbf{y}; \theta)}{f(\mathbf{z}; \theta)} = \frac{g'(T'(\mathbf{y}); \theta)h'(\mathbf{y})}{g'(T'(\mathbf{z}); \theta)h'(\mathbf{z})} = \frac{h'(\mathbf{y})}{h'(\mathbf{z})}.$$

Since ratio does not depend on θ , we have by assumption $T(\mathbf{y}) = T(\mathbf{z})$. Hence T is a function of T' ; so is minimal by arbitrary choice of T' . \square

Example (Bernoulli Trials)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \text{Bernoulli}(\theta)$. Let $\mathbf{z}, \mathbf{y} \in \{0, 1\}^n$ be two possible outcomes. Then

$$\frac{f(\mathbf{z}; \theta)}{f(\mathbf{y}; \theta)} = \frac{\theta^{\sum z_i} (1 - \theta)^{n - \sum z_i}}{\theta^{\sum y_i} (1 - \theta)^{n - \sum y_i}}$$

which is constant if and only if $T(\mathbf{z}) = \sum z_i = \sum y_i = T(\mathbf{y})$, so that T is minimally sufficient.

Example (Minimal sufficiency for k -parameter exponential families)

An i.i.d. sample $(Y_1, \dots, Y_n)^\top$ from an exponential family has joint distribution

$$f_{Y_1, \dots, Y_n}(y_1, \dots, y_n) = \exp \left\{ \sum_{j=1}^k \phi_j \tau_j(y_1, \dots, y_n) - n\gamma(\phi_1, \dots, \phi_k) + \sum_{i=1}^n S(y_i) \right\}$$

where $\tau_j(y_1, \dots, y_n) = \sum_{i=1}^n T_j(y_i)$, as before. If the $\{T_j\}_{j=1}^k$ are non-trivial, the ratio $f(\mathbf{y})/f(\mathbf{z})$ will be constant with respect to (ϕ_1, \dots, ϕ_k) if and only if as (ϕ_1, \dots, ϕ_k) varies, the quantity below remains constant.

$$\sum_{j=1}^k \phi_j (\tau_j(y_1, \dots, y_n) - \tau_j(z_1, \dots, z_n))$$

So if (ϕ_1, \dots, ϕ_k) range over an open parameter space of dimension k , this must imply that

$$\tau_j(y_1, \dots, y_n) = \tau_j(z_1, \dots, z_n).$$

Conversely, when the latter is true, the density ratio is clearly independent of the parameters, and so the statistic $\tau(\mathbf{y}) = (\tau_1(\mathbf{y}), \dots, \tau_k(\mathbf{y}))$ is minimally sufficient for (ϕ_1, \dots, ϕ_k) .

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 $F(y_1, \dots, y_n; \theta)$?

We will now:

- (2a) Review important special cases where sampling distributions can be **exactly determined**, focussing on the iid sampling case.
 - ↪ focussing on sufficient statistics of Gaussian and exponential families.
- (2b) Study ways of getting **approximations to the sampling behaviour** when the precise form is not explicitly available or is tedious (**stochastic convergence**).

Theorem (Sampling Distribution of Gaussian Sufficient Statistics)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, and define

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i \quad \& \quad S^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

The pair (\bar{Y}, S^2) is minimally sufficient for (μ, σ^2) and:

- ❶ The sample mean is distributed as $\bar{Y} \sim N(\mu, \sigma^2/n)$.
- ❷ The random variables \bar{Y} and S^2 are independent.
- ❸ The random variable S^2 satisfies $\frac{n-1}{\sigma^2} S^2 \sim \chi_{n-1}^2$.

Corollary (Moments of Sufficient Statistics)

If $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$, then

$$\mathbb{E}[\bar{Y}] = \mu, \quad \text{var}(\bar{Y}) = \frac{\sigma^2}{n}, \quad \mathbb{E}[S^2] = \sigma^2, \quad \text{var}(S^2) = \frac{2\sigma^4}{n-1}.$$

(which is why we typically prefer factor of $(n-1)^{-1}$ instead of n^{-1} in S^2)

Theorem (Sums of Gaussian Squares)

Let $\{Z_1, \dots, Z_k\}$ be i.i.d. $N(0, 1)$ random variables. Then,

$$Z_1^2 + \dots + Z_k^2 \sim \chi_k^2.$$

Recall that a random variable X is said to follow the chi-square distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim \chi_k^2$, if it holds that $X \sim \text{Gamma}(k/2, 1/2)$. In other words,

$$f_X(x; k) = \begin{cases} \frac{1}{2^{k/2} \Gamma(\frac{k}{2})} x^{\frac{k}{2}-1} e^{-\frac{x}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance and moment generating function of $X \sim \chi_k^2$ are given by

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

Theorem (Student's Statistic and its Sampling Distribution)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} N(\mu, \sigma^2)$. Then, the empirically standardised mean satisfies

$$\frac{\bar{Y} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

Recall that a random variable X is said to follow Student's t distribution with parameter $k \in \mathbb{N}$ (called the number of degrees of freedom), denoted $X \sim t_k$, if,

$$f_X(x; k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right) \sqrt{k\pi}} \left(1 + \frac{x^2}{k}\right)^{-\frac{k+1}{2}},$$

Assuming $k > 2$, the mean and variance of $X \sim t_k$ are given by

$$\mathbb{E}[X] = 0, \quad \text{var}[X] = \frac{k}{k-2}.$$

The mean is undefined for $k = 1$ and the variance is undefined for $k \leq 2$. The moment generating function is undefined for any $k \in \mathbb{N}$.

Theorem (Ratios of Gaussian Sums of Squares and F-Statistic)

Let $Y_1 \sim \chi_{d_1}^2$ and $Y_2 \sim \chi_{d_2}^2$ be independent random variables. Then,

$$\frac{Y_1/d_1}{Y_2/d_2} \sim F_{d_1, d_2}.$$

A random variable X is said to follow the Snedecor-Fisher F distribution with parameters $d_1 \in \mathbb{N}$ and $d_2 \in \mathbb{N}$, denoted $X \sim F_{d_1, d_2}$, if

$$f_X(x; d_1, d_2) = \begin{cases} \frac{1}{B(\frac{d_1}{2}, \frac{d_2}{2})} \left(\frac{d_1}{d_2}\right)^{d_1/2} x^{\frac{d_1}{2}-1} \left(1 + \frac{d_1}{d_2}x\right)^{-\frac{d_1+d_2}{2}}, & \text{if } x \geq 0 \\ 0 & \text{if } x < 0. \end{cases}$$

The mean, variance of $X \sim F_{d_1, d_2}$ are given by

$$\mathbb{E}[X] = \frac{d_2}{d_2 - 2}, \text{ provided } d_2 > 2, \text{ var}[X] = \frac{2d_2^2(d_1 + d_2 - 2)}{d_1(d_2 - 4)(d_2 - 2)^2} \text{ provided } d_2 > 4.$$

The moment generating function does not exist.

Some remarks:

- Notice that in all cases we considered sampling distributions when the sample is iid Normal.
- It's possible to consider samples that are jointly Normally distributed but not iid. This we postpone till later, when we bring in covariates and study regression.
- It's not particularly productive to insist on the formulae for the densities of χ^2 , t and F distributions.
 - Much better to think of them as being implicitly defined via their relation to iid Gaussian sampling (sums of squares and their normalised ratios, etc).
 - This is what is crucial to remember, along with the relation of their parameters (degrees of freedom) to the setting at hand.

Theorem (Sampling from an Exponential Family)

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} f$, where

$$f(y) = \exp \left\{ \sum_{i=1}^k \phi_i T_i(y) - \gamma(\phi_1, \dots, \phi_k) + S(y) \right\}, \quad \phi = (\phi_1, \dots, \phi_k)^T \in \Phi \subseteq \mathbb{R}^k$$

be a density of a k -parameter exponential family form.

If Φ is open, then:

- ❶ The minimally sufficient statistic for ϕ is $\tau = (\tau_1, \dots, \tau_k)$ where

$$\tau_j(y_1, \dots, y_n) = \sum_{i=1}^n T_j(y_i).$$

- ❷ The function γ is infinitely differentiable in all k of its variables, and

$$\mathbb{E}[\tau] = n \nabla_{\phi} \gamma(\phi) \quad \text{and} \quad \text{cov}[\tau] = n \nabla_{\phi}^2 \gamma(\phi),$$

that is,

$$\mathbb{E}[\tau_j] = n \frac{\partial}{\partial \phi_j} \gamma(\phi_1, \dots, \phi_k) \quad \text{and} \quad \text{cov}\{\tau_m, \tau_j\} = n \frac{\partial^2}{\partial \phi_m \partial \phi_j} \gamma(\phi_1, \dots, \phi_k).$$

(recall that minimal sufficiency was already shown in an example)

Proof. (*)

Focus on case $k = 1$, so that $\tau_j = \tau = \sum_{i=1}^n T(Y_i)$. Let $\phi_0 \in \Phi$. Since Φ is open, there exists s sufficiently small so that $\phi_0 + s \in \Phi$. Now note that the MGF $M_{T(Y_1)}(u) = \mathbb{E}[\exp(uT(Y_1))]$ evaluated at s is

$$\int_{\mathbb{R}} e^{sT(y)} e^{\phi_0 T(y) - \gamma(\phi_0) + S(y)} dy = e^{\gamma(\phi_0 + s) - \gamma(\phi_0)} \underbrace{\int_{\mathbb{R}} e^{(\phi_0 + s)T(y) - \gamma(\phi_0 + s) + S(y)} dy}_{=1}$$

- ❶ Therefore $M_{T(Y_1)}(s) < \infty$ for s sufficiently small, and thus:
 - all moments of $T(Y_1)$ exist,
 - and $M_{T(Y_1)}(s)$ is infinitely differentiable on an open neighbourhood of 0.
- ❷ Furthermore, $\gamma(s + \phi_0)$ is infinitely differentiable for s small enough, i.e. γ is infinitely differentiable in an open neighbourhood of ϕ_0 . But ϕ_0 is arbitrary so γ is infinitely differentiable everywhere on Φ .

Now we we may differentiate w.r.t. s , and, setting $s = 0$, we get

$$\mathbb{E}[T(Y_1)] = \gamma'(\phi) \text{ and } \text{var}[T(Y_1)] = \gamma''(\phi).$$

The conclusion follows by the fact that $\tau = \sum_{i=1}^n T(Y_i)$. □

Unfortunately, the sampling distribution of a statistic $T(Y_1, \dots, Y_n)$ **isn't always obtainable in a closed/convenient form**

- Even when T is the sufficient statistic in an exponential family, we may not have a nice workable form for the sampling distribution.
- In this case we know that the sampling distribution is again a k -parameter exponential family, but its form may be tedious to work with.

General strategy:

Approximate the sampling distribution $F_{T(Y_1, \dots, Y_n)}$ by a simpler distribution G

Of course we must make sense of what it means that “the distribution $F_{T(Y_1, \dots, Y_n)}$ is approximated by the distribution G ”.

- 1 We will view $F_{T(Y_1, \dots, Y_n)}$ as a sequence of functions indexed by sample size n .
- 2 Thus, “approximation by G ” will be understood as a form of convergence of F_n to G as $n \rightarrow \infty$.
- 3 En route, we will also discover **a stronger form of convergence**.

Definition (Convergence in Distribution (Weak Convergence))

Let $\{F_n\}_{n \geq 1}$ be a sequence of distribution functions and G be a distribution function on \mathbb{R} . We say that F_n converges in distribution (or weakly) to G , and write $F_n \xrightarrow{d} G$, whenever

$$F_n(y) \xrightarrow{n \rightarrow \infty} G(y),$$

for all y constituting continuity points of G (i.e. all y such that $\lim_{\varepsilon \rightarrow 0} G(y + \varepsilon) = G(y)$).

Example

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} \mathcal{U}[0, 1]$, $M_n = \max\{Y_1, \dots, Y_n\}$, and $Q_n = n(1 - M_n)$.

$$\mathbb{P}[Q_n \leq y] = \mathbb{P}[M_n \geq 1 - y/n] = 1 - \left(1 - \frac{y}{n}\right)^n \xrightarrow{n \rightarrow \infty} 1 - e^{-y}$$

for all $y \geq 0$. Hence $Q_n \xrightarrow{d} Q$, with $Q \sim \exp(1)$.

Comments:

- 1 Convergence in distribution \equiv pointwise convergence of distribution function, with the exception that it is **not necessary to have convergence at discontinuity points of the limit**.
- 2 When $F_n(y) = \mathbb{P}[Y_n \leq y]$ for a sequence of random variables $\{Y_n\}_{n \geq 1}$ and $G(y) = \mathbb{P}[Z \leq y]$ for another random variable Z , we will abuse notation and write

$$Y_n \xrightarrow{d} Z.$$

- 3 Our aim of approximating the sampling distribution now translates into finding a random variable Z whose distribution is explicitly known and such that

$$T(Y_1, \dots, Y_n) \xrightarrow{d} Z$$

A stronger notion of convergence is **convergence in probability**:

Definition

When a sequence of random variables $\{Y_n\}$ satisfies $\mathbb{P}[|Y_n - Y| > \epsilon] \xrightarrow{n \rightarrow \infty} 0$ for all $\epsilon > 0$ and a given random variable Y , we say that Y_n converges in probability to Y , and write $Y_n \xrightarrow{P} Y$.

Example

Let $U_1, \dots, U_n \stackrel{iid}{\sim} \mathcal{U}[0, 1]$ and $M_n = \max\{U_1, \dots, U_n\}$. Fix $\epsilon \in (0, 1)$. Then

$$\mathbb{P}[|M_n - 1| > \epsilon] = \mathbb{P}[M_n > 1 + \epsilon] + \mathbb{P}[M_n < 1 - \epsilon] = 0 + (1 - \epsilon)^n \xrightarrow{n \rightarrow \infty} 0.$$

Hence $M_n \xrightarrow{P} 1$ as $n \rightarrow \infty$.

Comments:

- Convergence in probability implies convergence in distribution.
- Convergence in distribution **does not** imply convergence in probability
 - Consider $Z \sim \mathcal{N}(0, 1)$, $-Z + \frac{1}{n} \xrightarrow{d} Z$ but $-Z + \frac{1}{n} \not\xrightarrow{p} Z$.
- “ \xrightarrow{d} ” relates *distribution functions*. It says the probabilistic behaviour of a sequence Y_n becomes more and more alike to that of the limit Y .
- “ \xrightarrow{p} ” relates *random variables*. It says that the actual realisations of Y_n can be progressively approximated with high probability by those of Y .
- Both notions of convergence are *metrizable*
 - i.e. there exist metrics on the space of all random variables that are compatible with the notion of convergence.
 - Hence can use things such as the triangle inequality etc.

Theorem

$$(a) \quad Y_n \xrightarrow{p} Y \implies Y_n \xrightarrow{d} Y$$

$$(b) \quad Y_n \xrightarrow{d} c \implies Y_n \xrightarrow{p} c, \quad c \in \mathbb{R}.$$

Proof

(a) Let y be a continuity point of F_Y and $\epsilon > 0$. Then,

$$\begin{aligned} \mathbb{P}[Y_n \leq y] &= \mathbb{P}[Y_n \leq y, |Y_n - Y| \leq \epsilon] + \mathbb{P}[Y_n \leq y, |Y_n - Y| > \epsilon] \\ &\leq \mathbb{P}[Y \leq y + \epsilon] + \mathbb{P}[|Y_n - Y| > \epsilon] \end{aligned}$$

since $\{Y \leq y + \epsilon\}$ contains $\{Y_n \leq y, |Y_n - Y| \leq \epsilon\}$. Similarly,

$$\begin{aligned} \mathbb{P}[Y \leq y - \epsilon] &= \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| \leq \epsilon] + \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| > \epsilon] \\ &\leq \mathbb{P}[Y_n \leq y] + \mathbb{P}[|Y_n - Y| > \epsilon] \end{aligned}$$

which yields

$$\mathbb{P}[Y \leq y - \epsilon] - \mathbb{P}[|Y_n - Y| > \epsilon] \leq \mathbb{P}[Y_n \leq y].$$

Combining the two inequalities and “sandwiching” yields (a).

(b) Let F be the distribution function of a constant r.v. c ,

$$F(y) = \mathbb{P}[c \leq y] = \begin{cases} 1 & \text{if } y \geq c, \\ 0 & \text{if } y < c. \end{cases}$$

$$\begin{aligned} \mathbb{P}[|Y_n - c| > \epsilon] &= \mathbb{P}[\{Y_n - c > \epsilon\} \cup \{c - Y_n > \epsilon\}] \\ &= \mathbb{P}[Y_n > c + \epsilon] + \mathbb{P}[Y_n < c - \epsilon] \\ &\leq 1 - \mathbb{P}[Y_n \leq c + \epsilon] + \mathbb{P}[Y_n \leq c - \epsilon] \\ &\xrightarrow{n \rightarrow \infty} 1 - \underbrace{F(c + \epsilon)}_{\geq c} + \underbrace{F(c - \epsilon)}_{< c} = 0 \end{aligned}$$

Since $Y_n \xrightarrow{d} c$.



Now we explore the **stability of stochastic convergence notions under transformation**.

Theorem (Continuous Mapping Theorem)

Let $g : \mathbb{R} \rightarrow \mathbb{R}$ be a continuous on the range of Y . Then,

$$(a) \quad Y_n \xrightarrow{P} Y \implies g(Y_n) \xrightarrow{P} g(Y)$$

$$(b) \quad Y_n \xrightarrow{d} Y \implies g(Y_n) \xrightarrow{d} g(Y)$$

Theorem (Slutsky's Theorem)

Let $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$. Then

$$(a) \quad X_n + Y_n \xrightarrow{d} X + c$$

$$(b) \quad X_n Y_n \xrightarrow{d} cX$$

Proof of Slutsky's Theorem.

(a) We may assume $c = 0$. Let x be a continuity point of F_X . We have

$$\begin{aligned}\mathbb{P}[X_n + Y_n \leq x] &= \mathbb{P}[X_n + Y_n \leq x, |Y_n| \leq \epsilon] + \mathbb{P}[X_n + Y_n \leq x, |Y_n| > \epsilon] \\ &\leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]\end{aligned}$$

Similarly, $\mathbb{P}[X_n \leq x - \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] + \mathbb{P}[|Y_n| > \epsilon]$, therefore,

$$\mathbb{P}[X_n \leq x - \epsilon] - \mathbb{P}[|Y_n| > \epsilon] \leq \mathbb{P}[X_n + Y_n \leq x] \leq \mathbb{P}[X_n \leq x + \epsilon] + \mathbb{P}[|Y_n| > \epsilon]$$

Since ϵ is arbitrary, this proves (a) by taking $n \rightarrow \infty$.

(b) It suffices to assume $c = 0$ (since $(Y_n + c)X_n = X_n Y_n + X_n c$, so if we can show $X_n Y_n \xrightarrow{d} 0$, then (a) gives conclusion). Let $\epsilon, M > 0$:

$$\begin{aligned}\mathbb{P}[|X_n Y_n| > \epsilon] &\leq \mathbb{P}[|X_n Y_n| > \epsilon, |Y_n| \leq 1/M] + \mathbb{P}[|Y_n| \geq 1/M] \\ &\leq \mathbb{P}[|X_n| > \epsilon M] + \mathbb{P}[|Y_n| \geq 1/M] \\ &\xrightarrow{n \rightarrow \infty} \mathbb{P}[|X| > \epsilon M] + 0\end{aligned}$$

The first term can be made arbitrarily small by letting $M \rightarrow \infty$.



Theorem (General Version of Slutsky's Theorem)

Let $g : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$ be continuous and suppose that $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c \in \mathbb{R}$. Then, $g(X_n, Y_n) \rightarrow g(X, c)$ as $n \rightarrow \infty$.

↪ Notice that the general version of Slutsky's theorem **does not follow immediately** from the continuous mapping theorem.

- The continuous mapping theorem would be applicable if (X_n, Y_n) weakly converged **jointly** (i.e. their joint distribution) to (X, c) .
- But here we assume only **marginal convergence** (i.e. $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ separately, but their joint behaviour is unspecified).
- The key of the proof is that in the special case where $Y_n \xrightarrow{d} c$ where c is a constant, then **marginal convergence \iff joint convergence**.
- However if $X_n \xrightarrow{d} X$ where X is non-degenerate, and $Y_n \xrightarrow{d} Y$ where Y is non-degenerate, then the theorem **fails**.
- Notice that even the special cases (addition and multiplication) of Slutsky's theorem fail if both X and Y are non-degenerate.

We will later consider **joint stochastic convergence**.

Continuous mappings and Slutsky's lemma allow us to get new approximations from old ones.

→ But how do we get limit theorems in the first place?

Typically these stem from a clever use of the following two fundamental theorems:

Theorem (Law of Large Numbers)

Let $\{Y_n\}$ be independent random variables with $\mathbb{E}[Y_k] = \mu$ and $\mathbb{E}|Y_k| < \infty$, for all $k \geq 1$. Then, $n^{-1}(Y_1 + \dots + Y_n) \xrightarrow{P} \mu$.

Theorem (Central Limit Theorem)

Let $\{Y_n\}$ be an i.i.d sequence with mean μ and variance $\sigma^2 < \infty$. Then,

$$\sqrt{n} \left(\frac{1}{n} \sum_{i=1}^n Y_i - \mu \right) \xrightarrow{d} N(0, \sigma^2).$$

Said differently, for large n , $\bar{Y} \approx N(\mu, \sigma^2/n)$ or $Y_1 + \dots + Y_n \approx N(n\mu, n\sigma^2)$.

The following theorem combines Slutsky's lemma and the continuous mapping theorem in order to allow us to **transform** central limit theorems:

Theorem (The Delta Method)

Let $Z_n := a_n(X_n - \theta) \xrightarrow{d} Z$ where $a_n, \theta \in \mathbb{R}$ for all n and $a_n \uparrow \infty$. Let $g(\cdot)$ be continuously differentiable at θ . Then, $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

Proof

Taylor expanding around θ gives:

$$g(X_n) = g(\theta) + g'(\theta_n^*)(X_n - \theta), \quad \theta_n^* \text{ between } X_n, \theta.$$

Thus $|\theta_n^* - \theta| < |X_n - \theta| = a_n^{-1} \cdot |a_n(X_n - \theta)| = a_n^{-1} Z_n \xrightarrow{p} 0$ [by Slutsky]

Therefore, $\theta_n^* \xrightarrow{p} \theta$. By the continuous mapping theorem $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$.

$$\begin{aligned} \text{Thus} \quad a_n(g(X_n) - g(\theta)) &= a_n(g(\theta) + g'(\theta_n^*)(X_n - \theta) - g(\theta)) \\ &= g'(\theta_n^*)a_n(X_n - \theta) \xrightarrow{d} g'(\theta)Z. \end{aligned}$$

The delta method also applies even when $g'(\theta)$ is not continuous (proof harder).

We can apply this machinery to get the following result for the sampling distribution of a sufficient statistic in a 1-parameter exponential family:

Corollary

Let $Y_1, \dots, Y_n \stackrel{iid}{\sim} f$, where

$$f(x) = \exp \{ \phi T(x) - \gamma(\phi) + S(x) \}, \quad x \in \mathcal{X}$$

with $\phi \in \Phi \subseteq \mathbb{R}$ and

$$\overline{T}_n = \frac{1}{n} \sum_{i=1}^n T(X_i) = n^{-1} \tau(X_1, \dots, X_n).$$

If Φ is open, then γ is infinitely differentiable, and so

$$\sqrt{n}(\overline{T}_n - \gamma'(\phi)) \xrightarrow{d} N(0, \gamma''(\phi)).$$

The following more general CLT is often useful:

Theorem (Weighted Sum CLT)

Let $\{W_n\}$ be an i.i.d sequence of real random variables, with common mean 0 and variance 1. Let $\{\gamma_n\}$ be a sequence of real constants. Then, if

$$\sup_{1 \leq j \leq n} \frac{\gamma_j^2}{\sum_{i=1}^n \gamma_i^2} \xrightarrow{n \rightarrow \infty} 0 \implies \frac{1}{\sqrt{\sum_{i=1}^n \gamma_i^2}} \sum_{i=1}^n \gamma_i W_i \xrightarrow{d} N(0, 1).$$

- Supremum condition amounts to saying that, in the limit, any single component contributes a negligible proportion of the total variance.
- Coefficient sequence $\{\gamma_n\}$ might very well diverge, without contradicting the negligibility condition.

To have **joint convergence**, we need to consider random vectors:

Definition

Let $\{\mathbf{Y}_n\}$ be a sequence of random vectors of \mathbb{R}^d , and \mathbf{Y} a random vector of \mathbb{R}^d with $\mathbf{Y}_n = (Y_n^{(1)}, \dots, Y_n^{(d)})^\top$ and $\mathbf{Y} = (Y^{(1)}, \dots, Y^{(d)})^\top$. Define the distribution functions $F_{\mathbf{Y}_n}(\mathbf{y}) = \mathbb{P}[Y_n^{(1)} \leq y^{(1)}, \dots, Y_n^{(d)} \leq y^{(d)}]$ and $F_{\mathbf{Y}}(\mathbf{y}) = \mathbb{P}[Y^{(1)} \leq y^{(1)}, \dots, Y^{(d)} \leq y^{(d)}]$, for $\mathbf{y} = (y^{(1)}, \dots, y^{(d)})^\top \in \mathbb{R}^d$. We say that \mathbf{Y}_n converges in distribution to \mathbf{Y} as $n \rightarrow \infty$ (and write $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$) if for every continuity point of $F_{\mathbf{Y}}$ we have

$$F_{\mathbf{Y}_n}(\mathbf{y}) \xrightarrow{n \rightarrow \infty} F_{\mathbf{Y}}(\mathbf{y}).$$

There is a link between univariate and multivariate weak convergence:

Theorem (Cramér-Wold Device)

Let $\{\mathbf{Y}_n\}$ be a sequence of random vectors of \mathbb{R}^d , and \mathbf{Y} a random vector of \mathbb{R}^d . Then,

$$\mathbf{Y}_n \xrightarrow{d} \mathbf{Y} \iff \mathbf{u}^\top \mathbf{Y}_n \xrightarrow{d} \mathbf{u}^\top \mathbf{Y}, \quad \forall \mathbf{u} \in \mathbb{R}^d.$$

- Continuous mapping theorem and Slutsky's lemma generalise to vector case.
- In either case, continuity is understood in the **multidimensional sense**:
 - ➊ **Continuous mapping**: If $g : \mathbb{R}^p \rightarrow \mathbb{R}^d$ is continuous on the range of \mathbf{U} , and if $\mathbf{U}_n \xrightarrow{d} \mathbf{U}$ in \mathbb{R}^p , then $g(\mathbf{U}_n) \xrightarrow{d} g(\mathbf{U})$ in \mathbb{R}^d .
 - ➋ **Slutsky**: If $g : \mathbb{R}^p \times \mathbb{R}^q \rightarrow \mathbb{R}^d$ is continuous, and if $\mathbf{U}_n \xrightarrow{d} \mathbf{U}$ in \mathbb{R}^p and $\mathbf{W}_n \xrightarrow{d} \mathbf{u}$ in \mathbb{R}^q , for some deterministic \mathbf{u} , then $g(\mathbf{U}_n, \mathbf{W}_n) \xrightarrow{d} g(\mathbf{U}, \mathbf{u})$.

Convergence in probability easily generalises to the vector case:

Definition

When a sequence of random vectors $\{\mathbf{Y}_n\}$ in \mathbb{R}^d satisfies $\mathbb{P}[\|\mathbf{Y}_n - \mathbf{Y}\| > \epsilon] \xrightarrow{n \rightarrow \infty} 0$ for all $\epsilon > 0$ and a given random d -vector \mathbf{Y} , we say that \mathbf{Y}_n converges in probability to \mathbf{Y} , and write $\mathbf{Y}_n \xrightarrow{P} \mathbf{Y}$.

Theorem (Multivariate Law of Large Numbers)

Let $\{\mathbf{Y}_n\}$ be iid random vectors with $\mathbb{E}[\mathbf{Y}_k] = \boldsymbol{\mu}$ and $\mathbb{E}\|\mathbf{Y}_k\| < \infty$, for all $k \geq 1$. Then,

$$\frac{1}{n} \sum_{k=1}^n \mathbf{Y}_k \xrightarrow{p} \boldsymbol{\mu}$$

Theorem (Multivariate CLT)

Let $\{\mathbf{X}_n\}$ be an iid sequence of random vectors in \mathbb{R}^d with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Omega}$ and define $\bar{\mathbf{X}}_n := \sum_{m=1}^n \mathbf{X}_m / n$. Then, $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{d} \mathbf{Z} \sim \mathcal{N}_d(0, \boldsymbol{\Omega})$ where $\mathbf{Y}_n \xrightarrow{d} \mathbf{Y}$ means $F_{\mathbf{Y}_n}(u) \rightarrow F_{\mathbf{Y}}(u)$ for any continuity point $u \in \mathbb{R}^d$ of $F_{\mathbf{Y}}$.

OK, but **how fast?**

Theorem (Berry-Essen)

In the same setting as the previous theorem, take $\boldsymbol{\mu} = 0$ and $\boldsymbol{\Omega} = I$, then

$$\sup_{\mathbf{u} \in \mathbb{R}^d} \left| F_{\sqrt{n}\bar{\mathbf{Y}}}(\mathbf{u}) - F_{\mathbf{Z}}(\mathbf{u}) \right| \leq C n^{-1/2} d^{1/4} \mathbb{E}\|\mathbf{Y}_i\|^3.$$

We also have a vector version of the Delta Method:

Theorem (Delta Method – vector case)

Let $\mathbf{Z}_n := a_n(\mathbf{X}_n - \mathbf{u}) \xrightarrow{d} \mathbf{Z}$ in \mathbb{R}^d where $a_n \in \mathbb{R}$, $\mathbf{u} \in \mathbb{R}^d$ and $a_n \uparrow \infty$. Let $g : \mathbb{R}^d \rightarrow \mathbb{R}^p$ be continuously differentiable at \mathbf{u} . Then,

$$a_n(g(\mathbf{X}_n) - g(\mathbf{u})) \xrightarrow{d} J_g(\mathbf{u})\mathbf{Z},$$

where $J_g(\mathbf{y})$ is the $p \times d$ Jacobian matrix of g ,

$$J_g(\mathbf{y}) = \begin{bmatrix} \frac{\partial}{\partial x_1} g_1(\mathbf{y}) & \cdots & \frac{\partial}{\partial x_d} g_1(\mathbf{y}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial x_1} g_p(\mathbf{y}) & \cdots & \frac{\partial}{\partial x_d} g_p(\mathbf{y}) \end{bmatrix}.$$