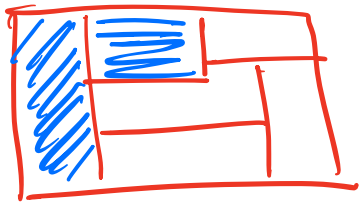
Statistics for Data Science: Week 3

Myrto Limnios and Rajita Chandak

Institute of Mathematics – EPFL

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

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



- $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} F_\theta$ and $T(\boldsymbol{Y})$ a statistic.

$T(y) = \sum y_i = t$, y

- The *level sets* or *contours* of T are the sets

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

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

$\hookrightarrow$ Clearly, T is constant when restricted to a level set.

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

- $\boldsymbol{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} F_\theta$ and $T(\boldsymbol{Y})$ a statistic.

- The *level sets* or *contours* of T are the sets

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

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

$\hookrightarrow$ Clearly, T is constant when restricted to a level set.

- Any realization of $\boldsymbol{Y}$ 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 $\boldsymbol{Y}$ conditional on a given level set A_t of T, $F_{\boldsymbol{Y}|T=t}(\boldsymbol{y})$...

- Suppose $F_{\boldsymbol{Y}|T=t}$ changes depending on θ: we are losing information.
- Suppose $F_{\boldsymbol{Y}|T=t}$ is functionally independent of θ

  $\Longrightarrow$ Then $\boldsymbol{Y}$ contains no information about θ on the set A_t

  $\Longrightarrow$ In other words, $\boldsymbol{Y}$ is ancillary for θ on A_t

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

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

Definition (Sufficient Statistic)

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

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

does *not* depend on θ.

Example (Coin Tossing)

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

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

Example (Coin Tossing)

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

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

- 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

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

Theorem (Fisher-Neyman Factorization Theorem)

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

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

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

Theorem (Fisher-Neyman Factorization Theorem)

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

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

Example

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

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

$g(T(y), \theta)$ $h(y)$

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

Proof of Neyman-Fisher Theorem - Discrete Real Statistic.

Suppose first that T is sufficient. Then

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

Since T is sufficient, $\mathbb{P}[\boldsymbol{Y}=\boldsymbol{y}|T=T(\boldsymbol{y})]$ is independent of θ and so $f(y;\theta) = g(T(\boldsymbol{y});\theta)h(\boldsymbol{y})$.

Proof of Neyman-Fisher Theorem - Discrete Real Statistic.

Suppose first that T is sufficient. Then

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

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

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

which does not depend on θ. □

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

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$. 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

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

Consequently, Fisher-Neyman factorisation implies that the statistic

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

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

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

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

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

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

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

where

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

So the statistic

$$\tau(Y_1, ..., Y_n) = \left(\tau_1(Y_1, ..., Y_n), ..., \tau_k(Y_1, ..., Y_n)\right)^\top$$

is sufficient for $(\phi_1, ..., \phi_k)$ 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 $T = T(\boldsymbol{Y})$ is said to be *minimally sufficient* for the parameter θ if it is sufficient for θ and for any other sufficient statistic $S = S(\boldsymbol{Y})$ there exists a function $g(\cdot)$ with

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

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

Proof.(*)

Assume for simplicity that $f(\boldsymbol{y}; \theta) > 0$ for all $\boldsymbol{y} \in \mathbb{R}^n$ and $\theta \in \Theta$. Let $\mathcal{T} = \{T(\boldsymbol{u}) : u \in \mathbb{R}^n\}$ 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 $\boldsymbol{w}_t \in A_t$. Notice that for any $\boldsymbol{y}$, $\boldsymbol{w}_{T(\boldsymbol{y})}$ is in the same level set as $\boldsymbol{y}$, so that

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

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

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

and sufficiency follows from the Fisher-Neyman factorization theorem.

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

$$\frac{f(\boldsymbol{y};\theta)}{f(\boldsymbol{z};\theta)} \overset{\text{defn.}}{=} \frac{g'(T'(\boldsymbol{y});\theta)h'(\boldsymbol{y})}{g'(T'(\boldsymbol{z});\theta)h'(\boldsymbol{z})} = \frac{h'(\boldsymbol{y})}{h'(\boldsymbol{z})}.$$

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

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

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

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

Example (Bernoulli Trials)

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

density

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

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

Example (Minimal sufficiency for k-parameter exponential families)

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

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

where $\tau_j(y_1, \ldots, y_n) = \sum_{i=1}^{n} T_j(y_i)$, as before. If the $\{T_j\}_{j=1}^{k}$ are non-trivial, the ratio $f(\boldsymbol{y})/f(\boldsymbol{z})$ will be constant with respect to $(\phi_1, \ldots, \phi_k)$ if and only if as $(\phi_1, \ldots, \phi_k)$ varies, the quantity below remains constant.

$$\sum_{j=1}^{k} \phi_j \big(\tau_j(y_1, \ldots, y_n) - \tau_j(z_1, \ldots, z_n)\big) = c \perp (\phi_1, \ldots \phi_k)$$

So if $(\phi_1, ..., \phi_k)$ range over an open parameter space of dimension k, this must imply that

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

Conversely, when the latter is true, the density ratio is clearly independent of the parameters, and so the statistic $\tau(\boldsymbol{y}) = (\tau_1(\boldsymbol{y}), ..., \tau_k(\boldsymbol{y}))$ is minimally sufficient for $(\phi_1, ..., \phi_k)$.

Anything we do will be a function $T(Y_1, \ldots, Y_n)$ 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 $F(y_1, ..., y_n; \theta)$?

$T_n = \frac{1}{n}\sum_{i=1}^{n} Y_i$ $\quad f_T$

Anything we do will be a function $T(Y_1, \ldots, Y_n)$ 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 $F(y_1, ..., y_n; \theta)$?

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 $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *and define*

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

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

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

degrees of freedom

Theorem (Sampling Distribution of Gaussian Sufficient Statistics)

Let $Y_1, ..., Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *and define*

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

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

1. *The sample mean is distributed as* $\bar{Y} \sim N(\mu, \sigma^2/n)$.
2. *The random variables* $\bar{Y}$ *and* S^2 *are independent.*
3. *The random variable* S^2 *satisfies* $\dfrac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

Corollary (Moments of Sufficient Statistics)

If $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2)$, *then*

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

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

Theorem (Sums of Gaussian Squares)

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

$$Z_1^2 + \ldots + Z_k^2 \sim \chi_k^2$$

Theorem (Sums of Gaussian Squares)

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

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

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 $X \sim Gamma(k/2, 1/2)$. In other words,

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

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

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

Theorem (Student's Statistic and its Sampling Distribution)

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

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

$\sqrt{n}\left(\frac{\bar{Y}-\mu}{\sigma}\right)$ ← standardized statistic

S ← unknown quantity

Theorem (Student's Statistic and its Sampling Distribution)

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

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

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 $X \sim \mathrm{t}_k$, 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 $X \sim \mathrm{t}_k$ are given by

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

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 $Y_1 \sim \chi^2_{d_1}$ *and* $Y_2 \sim \chi^2_{d_2}$ *be independent random variables. Then,*

$$\frac{Y_1/d_1}{Y_2/d_2} \sim \mathsf{F}_{d_1,d_2}.$$

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

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

$$\frac{Y_1/d_1}{Y_2/d_2} \sim \mathsf{F}_{d_1,d_2}.$$

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 $X \sim \mathsf{F}_{d_1,d_2}$, if

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

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

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

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.
  - $\hookrightarrow$ 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).
  - $\hookrightarrow$ 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 $Y_1, ..., Y_n \overset{iid}{\sim} f$, *where*

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

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

If Φ *is open, then:*

1. *The minimally sufficient statistic for* $\boldsymbol{\phi}$ *is* $\boldsymbol{\tau} = (\tau_1, ..., \tau_k)$ *where*
$$\tau_j(y_1, ..., y_n) = \textstyle\sum_{i=1}^{n} T_j(y_i).$$
2. *The function* γ *is infinitely differentiable in all* k *of its variables, and*
$$\mathbb{E}[\boldsymbol{\tau}] = n\nabla_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}) \quad \text{and} \quad \mathrm{cov}[\tau] = n\nabla^2_{\boldsymbol{\phi}}\gamma(\boldsymbol{\phi}),$$
that is,
$$\mathbb{E}[\tau_j] = n\frac{\partial}{\partial\phi_j}\gamma(\phi_1, ..., \phi_k) \quad \text{and} \quad \mathrm{cov}\{\tau_m, \tau_j\} = n\frac{\partial^2}{\partial\phi_m\partial\phi_j}\gamma(\phi_1, ..., \phi_k).$$

(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 $M_{T(Y_1)}(u) = \mathbb{E}[\exp(sT(Y_1))]$ evaluated at s is

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

1. 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.
2. 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

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

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

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

- $\hookrightarrow$ Even when T is the sufficient statistic in an exponential family, we may not have a nice workable form for the sampling distribution.
- $\hookrightarrow$ 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.

Unfortunately, the sampling distribution of a statistic $T(Y_1, \ldots, Y_n)$ 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 $F_{T(Y_1,\ldots,Y_n)}$ by a simpler distribution G

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

1. We will view $F_{T(Y_1,\ldots,Y_n)}$ 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 $n \to \infty$. $F_n \longrightarrow G = F_\infty$
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

$$F_n(y) \overset{n\to\infty}{\longrightarrow} G(y),$$

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

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

$$F_n(y) \overset{n\to\infty}{\longrightarrow} G(y),$$

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

Example

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

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

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.

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

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

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

A stronger notion of convergence is convergence in probability:

Definition

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

A stronger notion of convergence is convergence in probability:

Definition

When a sequence of random variables $\{Y_n\}$ satisfies $\mathbb{P}[|Y_n - Y| > \epsilon] \overset{n\to\infty}{\longrightarrow} 0$ 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

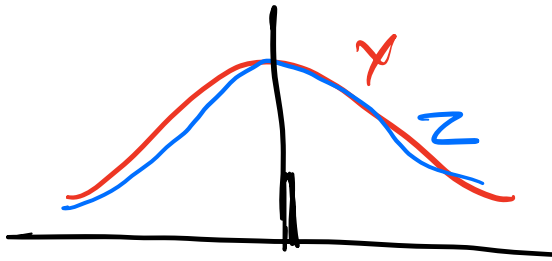
Let $U_1, \ldots, U_n \overset{iid}{\sim} \mathcal{U}[0,1]$ and $M_n = \max\{U_1, ..., U_n\}$. Fix $\epsilon \in (0,1)$. Then

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

Hence $M_n \xrightarrow{p} 1$ as $n \to \infty$.

Comments:

$X \xrightarrow{p} Y \Rightarrow X \xrightarrow{d} Y$



- 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.

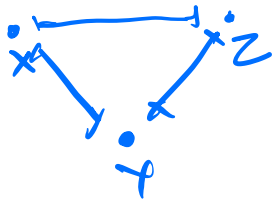
Metric $d(\cdot,\cdot)$ on some space $\mathcal{X} = \mathbb{R}^n$

① $d(x,x) = 0$

② $d(x,y) = d(y,x)$

③ $d(x,y) \geq 0$ $\quad d(x,y) = 0 \Leftrightarrow x = y$

④ "triangle inequality" $\quad d(x,z) \leq d(x,y) + d(y,z)$



$d(X,Y) = 0$ iff $X \xrightarrow[p]{d} Y$

Theorem

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

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

Theorem

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

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

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,

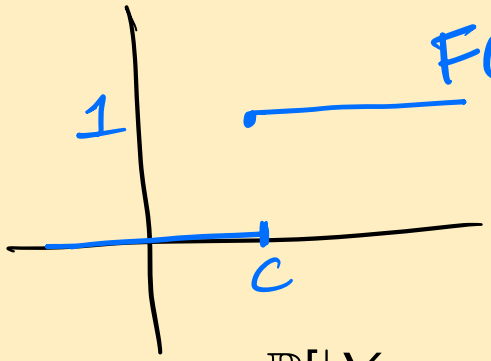
$$\begin{aligned} \mathbb{P}[Y \leq y - \epsilon] &= \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| \leq \epsilon] + \mathbb{P}[Y \leq y - \epsilon, |Y_n - Y| > \epsilon] \\ &\leq \mathbb{P}[Y_n \leq y] + \mathbb{P}[|Y_n - Y| > \epsilon] \end{aligned}$$

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,



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

splitting abs value

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

$F(c+\epsilon) \to 1$, $F(c-\epsilon) \to 0$

Since $Y_n \xrightarrow{d} c$. □

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

Theorem (Continuous Mapping Theorem)

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

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

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

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

Theorem (Continuous Mapping Theorem)

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

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

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

Theorem (Slutsky's Theorem)

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

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

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

$Y_n \xrightarrow{d} Y$ not true in general : $X_n + Y_n \xrightarrow{d} X + Y$, $X_n Y_n \xrightarrow{d} XY$

Proof of Slutsky's Theorem.

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

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

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

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

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

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 $n \to \infty$.

(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 $\epsilon, M > 0$:

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

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

□

Theorem (General Version of Slutsky's Theorem)

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

Theorem (General Version of Slutsky's Theorem)

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

$\hookrightarrow$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 of 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.

$\hookrightarrow$ But how do we get limit theorems in the first place?

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

$\hookrightarrow$ 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, $n^{-1}(Y_1 + \ldots + Y_n) \xrightarrow{p} \mu$.

$= \frac{1}{n} \sum_{i=1}^{n} Y_i \xrightarrow{p} \mu$

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

$\hookrightarrow$ 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, $n^{-1}(Y_1 + \ldots + Y_n) \xrightarrow{p} \mu$.

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 $Y_1 + \ldots + Y_n \approx N(n\mu, n\sigma^2)$.

The following theorem combines Slutksy'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,* $a_n(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Z$.

The following theorem combines Slutksy'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:

first-order

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

$+ \frac{g''(\theta)(X_n-\theta)^2}{2!}$

multiply & divide by a_n

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 $g'(\theta_n^*) \xrightarrow{p} g'(\theta)$.

taylor exp.

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

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 $Y_1, \ldots, Y_n \overset{iid}{\sim} f$, *where*

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

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

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

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

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

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

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

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

$$F_{\boldsymbol{Y}_n}(\boldsymbol{y}) \overset{n\to\infty}{\longrightarrow} F_{\boldsymbol{Y}}(\boldsymbol{y}).$$

There is a link between univariate and multivariate weak convergence:

Theorem (Cramér-Wold Device)

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

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

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

Convergence in probability easily generalises to the vector case:

Definition

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

Theorem (Multivariate Law of Large Numbers)

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

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

Theorem (Multivariate CLT)

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

Theorem (Multivariate Law of Large Numbers)

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

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

Theorem (Multivariate CLT)

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

OK, but how fast?

$I_d = \begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix}$ $\sqrt{n}\left(\frac{1}{n}\sum Y_i - \mu\right) \xrightarrow{p} 0$

Theorem (Berry-Esseen)

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

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

$\frac{\mathbb{E}[\|Y_i\|^3]}{\sqrt{n}}$

rate of convergence

We also have a vector version of the Delta Method:

Theorem (Delta Method – vector case)

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

$$a_n(g(X_n) - g(\boldsymbol{u})) \xrightarrow{d} J_g(\boldsymbol{u})\boldsymbol{Z},$$

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

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

Point Estimation

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.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

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.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The first sort of assertion we wish to make is:

1. **Point Estimation**. Given realisation $(Y_1, \ldots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

$T(Y) = \hat{\theta} \xrightarrow{P} \theta$

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.
4. Use the realisation $\{Y_1, \ldots, Y_n\}$ in order to make assertions concerning the true value of θ, and quantify the uncertainty associated with these assertions.

The first sort of assertion we wish to make is:

1. **Point Estimation**. Given realisation $(Y_1, \ldots, Y_n)^\top$ from $F(y_1, ..., y_n; \theta)$, how can we produce an educated guess for the unknown true parameter θ?

How? With a point estimator!

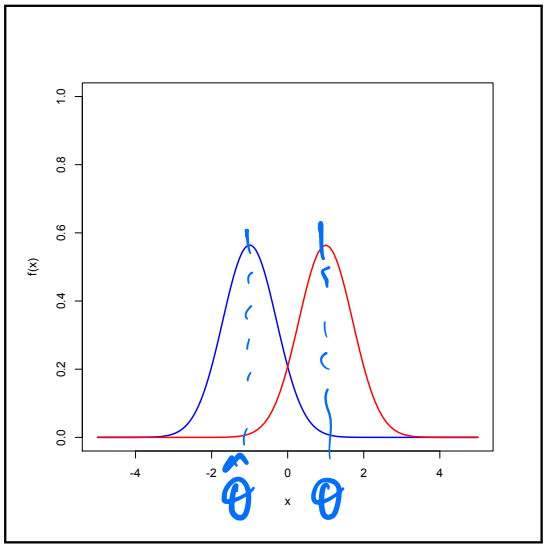
Definition (Point Estimator)

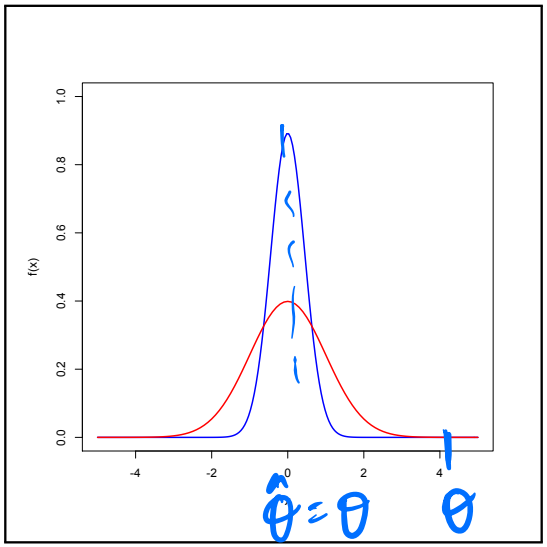
A statistic with codomain Θ is called a *point estimator*, i.e. a point estimator is a statistic $T : \mathcal{Y}^n \to \Theta$.

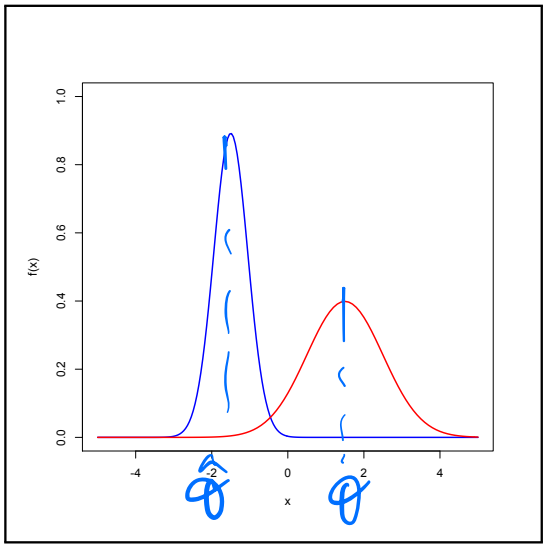
Since the objective of an estimator is to estimate the θ that generated the data, we typically denote it by $\hat{\theta}(Y_1, ..., Y_n)$, or just $\hat{\theta}$. Note that θ is a deterministic parameter, whereas $\hat{\theta}$ is a random variable.

But which estimator?

- Any statistic taking values in Θ could be used!
- Simpler yet, if we are given some $\hat{\theta}$, how do we judge its quality?
- Since estimators are *random variables*, every different realisation of the sample $(Y_1, \ldots, Y_n)$ will produce a different realised value for $\hat{\theta}$.
- A good estimator should be such that it typically manifests realisations that fall near the true θ.
- More precisely, the sampling distribution of an estimator should be concentrated around the true parameter value θ.







There is a multitude of criteria one can use, but a typical choice is to focus on two basic notions of location and spread for $\hat{\theta}$: its mean and variance

Why?

1. Ease of interpretation. The expectation $\mathbb{E}[\hat{\theta}]$ informs us whether the sampling distribution is located near the truth, whereas the variance $\mathrm{var}[\hat{\theta}]$ quantifies the degree of concentration around the expectation.

2. Central limit theory. Using our theory of stochastic convergence, we can often approximate the sampling distribution of $\hat{\theta}$ by a normal distribution. The latter is fully described by its mean and variance.

3. Concentration inequalities. We can often bound quantities such as $\mathbb{P}\{\|\hat{\theta} - \theta\| > \epsilon\}$ by means of moments.

A measure of precision that captures both mean and variance simultaneously is the mean squared error.

F_X, F_Y

$N(\mu, \sigma^2) = F_{(\mu,\sigma^2)}$

Definition (Mean Squared Error)

Let $\hat{\theta}$ be an estimator of a parameter θ corresponding to a model $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. The mean squared error of $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right].$$

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

Definition (Mean Squared Error)

Let $\hat{\theta}$ be an estimator of a parameter θ corresponding to a model $\{F_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$. The mean squared error of $\hat{\theta}$ is defined as

$$\mathrm{MSE}(\hat{\theta}, \theta) = \mathbb{E}\left[\left\|\hat{\theta} - \theta\right\|^2\right].$$

And here's the relation to means and variances:

Lemma (Bias-Variance Decomposition)

The MSE admits the decomposition

mean $\hat{\theta}$

$$\mathrm{MSE}(\hat{\theta}, \theta) = \underbrace{\left\|\mathbb{E}[\hat{\theta}] - \theta\right\|^2}_{\text{bias}^2} + \underbrace{\mathbb{E}\left[\left\|\hat{\theta} - \mathbb{E}(\hat{\theta})\right\|^2\right]}_{\text{variance}}.$$

$\mathrm{Var}(\hat{\theta})$