Theory and Methods of Inference

Daniele Zago

March 11, 2022

Contents

Preliminaries

In this lecture we will discuss some introductory topics related to probability theory, which comprise a fundamental tool to be used in statistical analysis. In general, statistical theory focuses on the interplay between probability as representing physical variability and encapsulating some aspects of epistemic uncertainty. We are interested in using probability models even though we do not have any kind of "experiment", insofar as they might produce similar data to what we have observed. The second tool of statistics is *quantifying the uncertainty* of the inferential procedures.

Suggested readings **reid2015** on the role of probability.
breiman2001a on the two cultures of statistics.
efron2020 on different approaches in statistics.

0.1 Empirical distribution function

References **vandervaart1998**

Empirical distribution functions are relevant when we want to use ***empirical distribution statistics***, i.e. when we want to compare a fitted cumulative distribution function to the observed empirical distribution function.

The ***bootstrap*** is closely related to the empirical distribution function: assume $Y_1, \dots, Y_n \overset{\text{iid}}{\sim} F(\cdot)$ and a statistic $T(Y_1, \dots, Y_n)$. The goal is to evaluate the distribution function $F_T(\cdot)$ of T, which usually is a function of F,

$$F_T(t; F).$$

The bootstrap comes into play by replacing the unknown F with its estimate given by the empirical distribution function,

$$F_T(t; \widehat{F}_n),$$

and simulation comes into place using the *plug-in principle*, i.e. when substituting the unknown F with an estimator $\widehat{F}$

$$\widehat{F}_T = F_T(t; \widehat{F}),$$

and sampling from $\widehat{F}_n$ or from $F(\cdot; \widehat{\vartheta})$ for numerical approximation.

Def. (Empirical distribution function)
We define the ***Empirical distribution function*** as the sample counterpart of $F_0(u) = \mathbb{P}(Y \leq u)$, given by

$$\widehat{F}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty, u]}(y_i).$$

Basic properties

› In general, we have that $n\widehat{F}_n(u) \sim \text{Bin}(n, F_0(u))$, and therefore

$$\mathbb{E}[\widehat{F}_n(u)] = F_0(u)$$

$$\mathbb{V}[\widehat{F}_n(u)] = \frac{1}{n}F_0(u)\big(1 - F_0(u)\big),$$

and moreover we have that if $u \neq v$,

$$\text{Cov}\left(\widehat{F}_n(u), \widehat{F}_n(v)\right) = \frac{1}{n}\Big(\min\{F_0(u), F_0(v)\} - F_0(u)F_0(v)\Big).$$

› By the strong law of large numbers, $\widehat{F}_n(y) \xrightarrow{\text{a.s.}} F_0(u)$ for every fixed u as $n \to \infty$.

Theorem 1 (Glivenko-Cantelli)
Let $D_n = \sup_{u\in\mathbb{R}} |\widehat{F}_n(u) - F_0(u)|$, *then* $\mathbb{P}(\lim_{n\to\infty} D_n = 0) = 1$.

Remark The above theorem guarantees uniform convergence of the distribution function, which is stronger than pointwise almost-sure convergence.

Remark Furthermore, if $F_0(u)$ is continuous then

$$\sqrt{n}\Big(\widehat{F}_n(u_1) - F_0(u_1),\ \ldots,\ \widehat{F}_n(u_k) - F_0(u_k)\Big) \xrightarrow{d} \mathcal{N}_k(0, \Sigma),$$

where $\Sigma = (\sigma_{ij})$ and

$$\sigma_{ij} = \text{Cov}\left(\widehat{F}_n(u_i), \widehat{F}_n(u_j)\right) = \frac{1}{n}\Big(\min\{F_0(u_i), F_0(u_j)\} - F_0(u_i)F_0(u_j)\Big).$$

0.2 Convergence of sums of r.v.'s

References **serfling1980**
vandervaart1998
billingsley2012

We will mostly use the following notions of convergence of random variables, namely:

Def. (convergence in probability)
X is said to ***convergence in probability*** to X if for any $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|x_n - X| \geq \varepsilon) = 0,$$

and we denote this convergence by $X_n \xrightarrow{P} X$.

Def. (convergence in distribution)

X with cumulative distribution function $F_{X_n}(x)$ is said to ***convergence in distribution*** to X with cumulative distribution function $F_X(x)$ if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x), \quad \text{for all } x \in \mathbb{R}^d,$$

such that x is a continuity point of F. We denote this convergence by $X_n \xrightarrow{d} X$.

Def. (Almost-sure convergence)

X_n is said to ***converge almost surely*** to X if

$$\mathbb{P}\Big(\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}\Big) = 1,$$

and we denote this convergence by $X_n \xrightarrow{d} X$.

Usage Convergence results are most useful when we cannot compute finite-sample distribution, since asymptotic results can be easier and allow us to to conduct approximate inference. Define $S_n = \sum_{i=1}^n Y_i$ as the sequence of sums and $\overline{Y}_n = S_n/n$ as the sequence of means.

Theorem 2 (Khintchine's weak LLN)

If $\{Y_i\}$ are sequences of i.i.d r.v.'s with $\mathbb{E}[Y_i] = \mu$, then $\overline{Y}_n \xrightarrow{P} \mu$.

Theorem 3 (Kolmogorov's strong LLN)

If $\{Y_i\}$ are sequences of i.i.d r.v.'s, then $\overline{Y}_n \xrightarrow{a.s.} \mu \iff \mathbb{E}[Y_i] = \mu < \infty$.

Theorem 4 (Lindberg-Lévy CLT)

Let $\{Y_i\}$ be a sequence of i.i.d r.v's with $\mathbb{E}[Y_i] = \mu$ and $\mathbb{V}[Y_i] = \sigma^2 < \infty$. Then,

$$\frac{\sqrt{n}(\overline{Y}_n.\mu)}{\sigma} \xrightarrow{d} N(0,1),$$

and its multivariate analogue with $\Sigma = \mathbb{E}[(Y_i - \mu)(Y_i - \mu)^\top]$,

$$\sqrt{n}(\overline{\boldsymbol{Y}}_n - \mu) \xrightarrow{d} N_d(0, \Sigma).$$

Theorem 5 (Chebyshev's weak LLN)

Let $\{Y_i\}$ be a sequence of independent random variables such that $\mathbb{E}[Y_i] = \mu_i < \infty$ and $\mathbb{V}[Y_i] = \sigma_i^2 < \infty$, then

$$\lim_{n\to\infty} \frac{1}{n^2} \sum_{i=1}^n \sigma_i^2 = 0 \implies \overline{Y}_n - \bar{\mu}_n \xrightarrow{P} 0,$$

where $\bar{\mu}_n = n^{-1} \sum_{i=1}^n \mu_i$.

Independence This theorem is also valid for sequences of uncorrelated random variables.

Theorem 6 (Kolmogorov's strong LLN)
Let $\{Y_i\}$ be a sequence of i.i.d r.v.'s with $\mathbb{E}[Y_i] = \mu_i$ and $\mathbb{V}[Y_i] = \sigma_i^2 < \infty$ for all i. If $\sum_{i=1}^{\infty} \sigma_i^2/i^2 < \infty$, then

$$\overline{Y}_n - \bar{\mu}_n \xrightarrow{a.s.} 0,$$

where $\bar{\mu}_n = n^{-1} \sum_{i=1}^{n} \mu_i$. The result is also valid for uncorrelated r.v.'s so long as they satisfy the more restrictive condition

$$\sum_{i=1}^{\infty} \frac{\sigma_i^2 \log^2 i}{i^2} < \infty.$$

Theorem 7 (Liapunov)
Let such that $\mathbb{E}[|Y_i - \mu_i|^3] = \beta_i < \infty$, and let

$$B_n = \left(\sum_{i=1}^{n} \beta_i \right)^{1/3}, \quad C_n = \left(\sum_{i=1}^{n} \sigma_i^2 \right)^{1/2}.$$

If $\lim_{n\to\infty} B_n/C_n = 0$, and ...

Theorem 8 (Lindberg-Feller)
The lindberg-Feller theorem ...

Remark The integral is the contribution of the tails to the variance of the distribution with respect to the whole distribution.

Theorem 9 (Slutsky)
Let $y_0 \in \mathbb{R}$ be constant, and let X, Y be random variables and $\{X_n\}$, $\{Y_n\}$ be sequences of random variables. Then,

$$X_n \xrightarrow{d} X \text{ and } Y_n \xrightarrow{P} y_0 \implies \begin{cases} X_n + Y_n \xrightarrow{d} X + y_0 \\ X_n Y_n \xrightarrow{d} X y_0 \end{cases}$$

then ...

Remark In general, we can apply usual limit operations when dealing with convergent sequences of random variables.

Theorem 10 (Convergence in probability)
let Y_n be a sequence of r.v.'s such that $Y_n \xrightarrow{P} c$. If $g(\cdot)$ is a continuous function on suppY, then

$$g(Y_n) \xrightarrow{P} g(c).$$

Theorem 11 (Convergence in distribution)
Let $\{Y_n\}$ be a sequence of r.v.'s such that

$$\sqrt{n}\big(Y_n - \vartheta\big) \xrightarrow{d} U,$$

with $\vartheta \in \mathbb{R}$ and U a non-degenerate function. Furthermore, let $\vartheta(\vartheta)$ be a continuously differentiable function with $g'(\vartheta) \neq 0$, then

$$\sqrt{n}\big(g(Y_n) - g(\vartheta)\big) \xrightarrow{d} g'(\vartheta)U.$$

0.3 Order statistics

References **arnold2008**

Def. (Order statistics)
The r^{th} ***order statistic*** is defined as the r^{th} value, $y_{(r)}$, of the ordered sample

$$y_{(1)} < y_{(2)} < \ldots < y_{(n-1)} < y_{(n)}.$$

Def. (Central order statistic)
The k_n^{th} order statistic is called ***central order statistic*** if $k_n/n \to c \in (0,1)$.

Remark The idea of a central order statistic is that the relative position of $X_{(n \cdot k_n)}$ stays roughly constant in the sample $X_1, \ldots, X_n$ as $n \to \infty$.

Lemma 1 (Uniform representation)
Let $Y_1, \ldots, Y_n$ be a sample from Y with cumulative distribution function F, then

$$U_1, \ldots, U_n = F(Y_1), \ldots, F(Y_n) \sim \mathit{Unif}(0,1).$$

Order statistics Since we can represent any distribution as $Y_i = F^{-1}(U_i)$, where $U_i \sim \text{Unif}(0,1)$, then by observing that F^{-1} is a monotone increasing function we can state that

$$\big(Y_{(1)}, \ldots, Y_{(n)}\big) \sim \big(F^{-1}(U_{(1)}), \ldots, U_{(n)}\big).$$

Theorem 12 (Exact distribution of $Y_{(r)}$)
The exact distribution of $Y_{(r)}$ has probability density function

$$p_{Y_{(r)}}(y) = \frac{n!}{(r-1)!(n-r)!} F(y)^{r-1} p(y) (1 - F(y))^{n-r},$$

Uniform In the special case $U \sim \text{Unif}(0,1)$, then we have an exact distribution for the r^{th} order statistic given by

$$U_{(r)} \sim \text{Beta}(r, n-r+1).$$

We can use this to approximate the expectation of the general order statistic $Y_{(r)}$ by using the approximate value

$$\mathbb{E}[U_{(r)}] = \frac{r}{n+1} \xrightarrow{1^{\text{st}} \text{ order}} \mathbb{E}[Y_{(r)}] \approx F^{-1}\left(\frac{r}{n+1}\right).$$

Theorem 13 (Joint distribution of pairs of order statistics)
We can write the joint distribution of pairs of order statistics as

$$p_{Y_{(r)},Y_{(s)}}(y,z) = \frac{n!}{(r-1)!(s-r-1)!(n-s)!} F(y)^{r-1} p(y) \big(F(z)-F(y)\big)^{s-r-1} p(z) \big(1-F(z)\big)^{n-s}.$$

Theorem 14 (Asymptotic distribution of order statistics)
let $r = [n\pi]$ with $0 < \pi < 1$, if $p(F^{-1}(\pi)) > 0$, then the asymptotic distributions of the r^{th} order statistic is given by

$$\sqrt{n} \frac{(Y_{(r)} - F^{-1}(\pi))}{\frac{\sqrt{\pi(1-\pi)}}{p(F^{-1}(\pi))}} \xrightarrow{d} Z \sim \mathcal{N}(0,1),$$

from which we conclude that

$$Y_{(r)} \dot{\sim} \mathcal{N}\Big(F^{-1}(\pi), \frac{\pi(1-\pi)}{n \cdot p\big(F^{-1}(\pi)\big)^2}\Big).$$

Proof.
Use the delta method on the asymptotic distribution of

$$U_{(r)} \dot{\sim} \mathcal{N}\Big(\pi, \frac{\pi(1-\pi)}{n}\Big),$$

with $Y_(r) = F^{-1}(U_{(r)})$.

□

Median The above theorem yields the normal approximation to the sample median, the minimum and the maximum.

0.3.1 Asymptotic distribution of extremes

Asymptotic distributions of extremes are especially relevant, since we can explicitly write the cumulative distribution function of $Y_{(n)}$ as

$$F_{Y_{(n)}}(y) = \mathbb{P}(y_1 \leq y, \ldots, y_n \leq y) = F(y)^n.$$

As $n \to \infty$, we have that $F_{Y_{(n)}}$ converges to the cumulative distribution function of a degenerate distribution with unit mass in the upper extreme of the support of $F(y)$.

Moreover, it is possible to identify suitable sequences of $\{a_n\}$ and $\{b_n\}$ with $b_n > 0$ such that

$$\mathbb{P}\big(b_n(Y_{(n)} - a_n) < y\big) = F(y/b_n + a_n)^n \xrightarrow{d} \Lambda(y), \tag{1}$$

which is the cumulative distribution function of a non-degenerate distribution. A famous result by **gnedenko1943** states that, if such limit exists, then it is one of the following three cumulative distribution functions:

$$\text{Fréchet:} \quad \Lambda_1 = e^{-y^{-\vartheta}} \mathbb{1}_{y>0}, \quad \vartheta > 0$$

$$\text{Weibull:} \quad \Lambda_2 = e^{-(-y)^{\vartheta}} \mathbb{1}_{y\leq 0} + \mathbb{1}_{y>0}, \quad \vartheta > 0$$

$$\text{Gumbel:} \quad \Lambda_3 = -e^{-y}$$

The previous result also apply to the sample minimum, since

$$Y_{(1)} = \min\{Y_1, \ldots, Y_n\} = -\max\{-Y_1, \ldots, -Y_n\}.$$

We can define ***domains of attraction*** in (1) for the distributions $\Lambda_1, \Lambda_2, \Lambda_3$ based on the behaviour of $F(y)$ on the upper tail of the distribution. Specifically,

1. $F(y)$ is in the domain of attraction of Λ_1 if the support of Y is unbounded and if

   $$1 - F(y) \sim cy^{-\vartheta}, \quad y \to +\infty, \vartheta > 0.$$

   The standardization is $(cn)^{-1/\vartheta} Y_{(n)}$.

2. $F(y)$ is in the domain of attraction of Λ_2 if $Y \leq y_0 < \infty$ and if

   $$1 - F(y) \sim c(y_0 - y)^{\vartheta}, \quad y \to \infty, \vartheta > 0,$$

   and the standardization needed is $(Y_{(n)} - y_0)(nc)^{1/\vartheta}$.

3. $F(y)$ is in the domain of attraction of Λ_3 if the support of Y has no finite upper bound and

   $$1 - F(y) \sim ce^{-y^h}, \quad y \longrightarrow \infty, h > 0,$$

   and the constants a_n and b_n are determined e.g. by

   $$a_n = F^{-1}(1 - 1/n),$$

   $$b_n = np_Y(a_n).$$

Example (Uniform distribution)
The uniform distribution is in the domain of attraction of Λ_2, and convergence to Λ_2 can be obtained by considering the sequence

$$n(Y_{(n)} - 1).$$

0.4 Density functions

References **billingsley2012**

Using the Radon-Nikodym theorem, there is a formal justification in the use of the same symbol for densities in both the discrete and continuous cases. In both cases, the integrals will be with respect to some dominating measure (e.g. counting measure in the discrete case and Lebesgue measure in the continuous case).

Theorem 15 (Radon-Nikodym)
Let $(\mathcal{Y}, \mathcal{B}, P)$ be a probability space and μ be a σ-finite measure such that $\mathbb{P} \ll \mu$. Then, there exists the ***density function*** *of $\mathbb{P}$ with respect to the dominating measure μ, $p = d\mathbb{P}/d\mu$ such that the probability of $B \in \mathcal{B}$ may be expressed as*

$$\mathbb{P}(B) = \int_B p \, \mathrm{d}\mu.$$

0.5 Moments and generating functions

References **pace1997**

Def. (r^{th} moment)
We define the ***r^{th} moment*** of a random variable Y as

$$\mu_r = \mathbb{E}[Y^r], \quad r = 1, 2, \ldots,$$

and the r^{th} ***central moment*** of Y as

$$\bar{\mu}_r = \mathbb{E}[(Y - \mu_1)^r], \quad r = 2, 3, \ldots$$

Def. (moment-generating function)
The ***moment-generating function*** of Y is defined as

$$M_Y(t) = \mathbb{E}[e^{tY}] = \int_{-\infty}^{\infty} e^{ty} p_Y(y) \, \mathrm{d}\mu, \quad t \in \mathbb{R}.$$

Remark $M_Y(t)$ takes positive values and, if it exists, is defined on an interval $(-\varepsilon, \varepsilon)$ containing the origin. It is not true that, in general, the moment-generating function is finite for all t in an open interval containing the origin.

Moments Whenever $M_Y(t)$ is finite in $(-\varepsilon, \varepsilon)$ for $\varepsilon > 0$, we can apply a Taylor series expansion around $t = 0$ to write

$$M_Y(t) = 1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \ldots,$$

from which we can calculate

$$\mu_r = \frac{\partial^r}{\partial t^r} M_Y(y)\Big|_{t=0}$$

Identification Whenever $M_Y(t)$ is finite in $(-\varepsilon, \varepsilon)$ for $\varepsilon > 0$, then both $M_Y(t)$ and the sequence of moments $\{\mu_r\}$ uniquely identifies the distribution.

Def. (Characteristic function)
The ***characteristic function*** of Y always exists and is defined as

$$C_Y(t) = \mathbb{E}[e^{itY}] = \int_{-\infty}^{\infty} e^{ity} p_Y(y)\, \mathrm{d}\mu.$$

Remark Since e^{ity} is bounded, the characteristic function is always finite and $C_Y(t)$ always characterizes the distribution.

There are inversion theorems (Fourier's inversion theorem), which yield the original density p from the characteristic function.

Theorem 16 (Fourier's inversion theorem)
Let $C_Y(t)$ be a characteristic function, then if C_Y is absolutely integrable, that is if

$$\int_{-\infty}^{\infty} |C_Y(t)|\, \mathrm{d}t < \infty,$$

then Y is absolutely continuous with density

$$p_Y(y) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ity} C_Y(t)\, \mathrm{d}t.$$

These functions are usually employed to describe convergence results, i.e. if Y_n has characteristic function $C_{Y_n}(t)$ and Y has characteristic function $C_Y(t)$, then

$$C_{Y_n}(t) \xrightarrow{n\to\infty} C_Y(t) \implies Y_n \xrightarrow{d} Y.$$

Characteristic functions are useful to assess limiting results, since for example if $S_n = \sum_{i=1}^n Y_i$ and the Y_i's are independent,

$$M_{S_n}(t) = \mathbb{E}[e^{t\sum_{i=1}^n Y_i}] = M_Y(t)^n,$$

$$C_{S_n}(t) = \mathbb{E}[e^{it\sum_{i=1}^n Y_i}] = C_Y(t)^n.$$

0.5.1 Infinitely-divisible and stable distributions

There are more general cases of convergence to a limit distribution even if the variance is not finite, specifically for a class of random variables called *stable distributions*.

Def. (Infinitely-divisible distribution)
If the characteristic function of a random variable Y can be represented in the form

$$C_Y(t) = \big(C_n(t)\big)^n, \quad \text{for all } n \in \mathbb{N}^+,$$

where $C_n(t)$ is a characteristic function, then the distribution of Y is said to be ***infinitely divisible***

Examples Some examples are the Poisson, normal and gamma distributions.

Remark If Y has a stable distribution, then it can be represented as a sum of n i.i.d random variables.

Def. (Stable distribution)
We say that Y has a ***stable distribution*** if given $Y_1, \dots, Y_n$ i.i.d and $S_n = \sum_{i=1}^n Y_i$, then there exist two sequences of constants a_n and $b_n > 0$ such that

$$Z = S_n/b_n - a_n \overset{\mathrm{d}}{=} Y.$$

Remark Y_i is said to be in the ***domain of attraction*** of the stable distribution of Z,

Remark The sequence b_n must have the form $b_n = n^{1/\vartheta} b_0(n)$ where $b_0(n)$ is a ***slowly-varying*** function on $[0, +\infty)$, i.e. for each $k > 0$,

$$\lim_{x\to\infty} \frac{b_0(kx)}{b_0(x)} = 1.$$

Example (Slowly-varying function)
For any $\beta \in \mathbb{R}$, the function $f(x) = \log^\beta x$ is slowly-varying, since for any $a > 0$,

$$\lim_{x\to\infty} \frac{f(ax)}{f(x)} = \lim_{x\to\infty} \frac{\log^\beta(ax)}{\log^\beta x} = \lim_{x\to\infty} \frac{\log^\beta x + \log^\beta a}{\log^\beta x}.$$

Exponent For every stable distribution we have $b_n = n^{1/\vartheta}$ for some ***characteristic exponent*** $0 \le \vartheta \le 2$.

Def. (Stable distribution (ii))

Equivalently, we can define the family of ***stable distributions*** as the subclass of infinitely-divisible distributions whose characteristic function is defined as

$$C_Y(t) = \exp\left\{ita - b|t|^\vartheta\left[1 + i\gamma \cdot (t/|t|)\omega(t,\vartheta)\right]\right\},$$

with $|\gamma| < 1$, $a \in \mathbb{R}$ $b \in \mathbb{R}^+$, and

$$\omega(t,\vartheta) = \begin{cases} \tan(\pi\vartheta/2) & \text{if } \vartheta \neq 1 \\ \frac{2}{\pi}\log|t| & \text{if } \vartheta = 1 \end{cases}$$

Remark The quantities a and b are location and scale parameters, while γ is an asymmetry parameter.

Remark If the distribution of Y is symmetric around zero, then $a = 0$ and

$$C_Y(t) = e^{-b|t|^\vartheta}.$$

Example (Cauchy)

The Cauchy distribution is stable with $\vartheta = 1$, and it is possible to see that

$$Y \sim \text{Cauchy}(0,1) \implies \overline{Y}_n \sim \text{Cauchy}(0,1).$$

Example (Normal)

The normal distribution is a stable distribution with $\vartheta = 2$.

Theorem 17 (Convergence to stable distribution)

Let $\{Y_i\}$ be a sequence of i.i.d r.v.'s with $\mathbb{V}[Y_i] = \infty$, then if we can standardize $Z_n = S_n/b_n - a_n$ such that

$$Z_n \xrightarrow{d} Z,$$

non degenerate, then the limit distribution is stable with characteristic exponent $0 < \vartheta < 2$.

This is of enormous interest in the econometric literature, since these distributions are appropriate for modelling stochastic processes which display heavy-tailed innovations.

Example (Characteristic function of Cauchy)

Because of independence, we can write

$$C_{\overline{Y}_n}(t) = \prod_{i=1}^n \mathbb{E}[e^{i\frac{t}{n}Y_i}] = \prod_{i=1}^n e^{-|t|/n} = e^{-|t|}.$$

Example (Lévy density)
The distribution

$$\dots$$

is also called a ***Lévy density*** and is a stable distribution.

0.5.2 Multivariate extensions

There are multivariate extensions to the moment-generating function and characteristic function, which require a somewhat different notation to simplify the expressions. Specifically, we use the notation

$$\mu^{i_1 \dots i_r} = \mathbb{E}[Y_{i_1} \cdots Y_{i_r}],$$

which allows us to write for example

$$\mu^i = \mathbb{E}[Y_i], \quad \mu^{ij} = \mathbb{E}[Y_i Y_j].$$

Thus, the moment-generating function of Y can be generalized as

$$M_Y(t) = \mathbb{E}[e^{t^\top Y}] = \mathbb{E}[e^{t_1 Y_1 + \dots + t_d Y_d}], \quad t \in \mathbb{R}^d,$$

and M_Y can be expanded in a multivariate power series with convergence radius $R \geq t_0$ as

$$M_Y(t) = 1 + \sum_{i=1}^{d} \mu^i t_i + \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \mu^{ij} t_i t_j + \dots \tag{2}$$

From (2), we can write the $i_1 \cdots i_r^{\text{th}}$ joint moment of Y as

$$\mu^{i_1 \dots i_r} = \frac{\partial^r M_Y(t)}{\partial t_{i_1} \cdots \partial t_{i_r}} \Big|_{t=0}.$$

If and only if the marginal components of Y are independently-distributed, then we can use the expected value factorization theorem to write

$$M_Y(t) = M_{Y_1}(t_1) \cdots M_{Y_d}(t_d).$$

Finally, we can also extend the cumulant-generating function to the multivariate case by setting

$$K_Y(t) = \log M_Y(t),$$

and by expanding it with a multivariate Taylor series analogously to (2), we obtain the joint cumulant of order r as

$$\kappa^{i_1 \dots i_r} = \frac{\partial K(t)}{\partial t_1 \cdots \partial t_r} \Big|_{t=0}.$$

Lecture 1: Statistical models

2022-03-03

In this lecture we review the fundamental tools of statistical modelling from a probabilistic point of view.

1.1 Fundamentals of statistical modelling

We have a fundamental assumption of y_i being a realization of a random vector – or more generally a *stochastic process* – Y having a partly unknown distribution. Dual role of probability

1. *Descriptive:* modelling data variability
2. *Epistemologic:* to evaluate uncertainty of our procedures

We assume the ***statistical model*** as a family $\mathcal{F}$ of probability distributions,

$$\mathcal{F} = \{p(y; \vartheta), \vartheta \in \Theta\},$$

which are at least qualitatively compatible with y. The family is indexed by the ***parameter*** $\vartheta \in \Theta$, and we say that

- $\mathcal{F}$ is ***correctly specified*** if $p^0(y) \in \mathcal{F}$ and ***misspecified*** otherwise;
- ϑ is ***identifiable*** if $p(y; \vartheta) \neq p(y; \vartheta')$ whenever $\vartheta \neq \vartheta'$.

In the case that the model $\mathcal{F}$ is both correctly specified and ϑ is identifiable, then $p^0(y) = p(y; \vartheta^0)$ for some $\vartheta^0 \in \Theta$.

A statistical model is called ***parametric*** if $\Theta \subseteq \mathbb{R}^p$ or $\Theta \equiv \mathbb{R}^p$. When $\vartheta = (\tau, h(\cdot))$ with $\tau \in T \subseteq \mathbb{R}^k$ we say that the model is ***semiparametric***. Finally, if $\vartheta = h(\cdot)$, the model is called ***nonparametric***.

Example (Semiparametric models)
Some examples of semiparametric models include

a) All densities of the form $p(y; \mu) = p_0(y - \mu)$.

b) Linear regression assuming $\mathbb{E}[Y_i] = \alpha + \beta x_i, \quad \mathbb{V}[Y_i] = \sigma^2$

c) Cox's proportional hazards model, as the set of distributions defined by

$$r_{Y_i}(y_i) = r_0(y_i) \exp\left\{\boldsymbol{x}_i^\top \beta\right\},$$

where $r_0(\cdot)$ is the ***baseline hazard function***.

The statistical model $\mathcal{F}$ can be defined in terms of any quantity of the distribution

- Distribution function $F(y) = \mathbb{P}(Y \geq y)$

› Moment-generating function $\mathbb{E}[e^{tY}]$

› Characteristic function $\mathbb{E}[e^{itY}]$

› Failure rate $r(y) = p(y)/\big(1 - F(y)\big)$, recall that

$$p(y) = r(y) \exp\Big\{ - \int_0^y r(t)\, \mathrm{d}t \Big\}$$

The specification of a particular model results as a consequence of some initial assumptions, e.g. normal or exponential results based on random sums and lack of memory, as well as asymptotic arguments for extreme value distributions, stable distributions, ...

There are some broad structures for statistical inference, namely Bayesian theory and frequentist inference, whose goals and quantities of interest may differ.

Bayesian approach

› *Conditional inference* by conditioning on the observed data y, and the object of analysis is the posterior distribution.

› *Equivalent* likelihood yield equivalent inferences, hence the sampling rule is ignored in Bayesian inference.

› *Subjectivist* view: probability describes a subject's state of knowledge, which is provided by the subject who analyzes the data.

› *Objective* view: using conventional of *default*/non-informative prior distributions, aiming at representing a prior open approach.

Frequentist approach

› *Repeated sampling*: inference based on y should be assessed by its behaviour in hypothetical repetitions of the same experiment that generated y, under the same initial conditions.

› Inference is a collection of *ad hoc* procedures to locate ϑ^0 in Θ.

› *Fisherian inference*: emphasis is on inference based on the likelihood function, significant tests and p-values. Uncertainty is evaluated through repeated sampling to be applied in a "relevant" sequence of replications of the data-generating process (conditioning on ancillary statistics).

› *Decision-frequentist inference*: a *loss function* is added to $\mathcal{F}$ instead of a prior distribution, and the idea is to find an inference procedure that minimizes the ***risk***, i.e. the expected loss for all $\vartheta \in \Theta$.

Example (Uniform distribution)
Since the uniform distribution $U(0, \vartheta)$ is a scale family, we can write

$$Y_{(n)} = \vartheta U_{(n)}, \quad U_{(n)} \sim \mathrm{Beta}(n, 1).$$

Using a decision-theoretic approach we obtain the optimal estimate

$$\widehat{\vartheta}(y) = y_{(n)}(n+1)/n.$$

Theorem 18 (Conjugate prior in an exponential model)
The conjugate prior distribution for a model

$$p(y|\varphi) = h(y)\exp\{\vartheta(\varphi)t(y) - G(\varphi)\}$$

is given by

$$\pi(\varphi|\nu,\chi) = \exp\{\nu\vartheta(\varphi) - \xi G(\varphi) + D(\nu,\xi)\},$$

which yields the posterior distribution for a random sample $y_1,\ldots,y_n$ *as*

$$\pi(\varphi|y) = \exp\Big\{\big(\nu + \sum_{i=1}^{n} t(y)\big)\vartheta(\varphi) - (\xi+n)G(\varphi) + D(\nu + \sum_{i=1}^{n} t(y), \xi+n\Big\}.$$

There are three classes of problems in statistical inference, namely

1. *Problems of specification*: the choice of the mathematical model $\mathcal{F}$ for Y.
2. *Problems of inference*: finding a suitable inferential procedure, i.e. the algorithm required to locate $p^0(y)$ within $\mathcal{F}$.
3. *Problems of distribution*: evaluating the accuracy of the estimation.

Def. (Statistic)
A ***statistic*** is a measurable function of the data $T(y)$.

Def. (Combinant)
A ***combinant*** is a function of both the data and the parameter, $q(y;\vartheta)$.

Def. (Induced model)
The ***induced model*** is the model induced by the statistic T assuming the data as random,

$$\mathcal{F}_T = \{p_T(t;\vartheta), \vartheta\in\Theta\}.$$

For a combinant,

$$\mathcal{F}_{Q_\vartheta} = \{p_{Q_\vartheta}(q;\vartheta,\vartheta'), \vartheta'\in\Theta\}.$$

Null distribution We say that the distribution of a combinant is the *null distribution* $q(y;\vartheta)$ when $\vartheta' = \vartheta$.

Some special cases of statistic and combinants are:

› ***Distribution-constant statistic***: the distribution of T is independent on ϑ in Θ.

Example (Distribution-constant statistic)
$Y_1, Y_2 \overset{\text{iid}}{\sim} N(\vartheta, 1)$, we have that $T = Y_1 - Y_2 \sim \mathcal{N}(0, 2)$ independently of ϑ.

› ***First-order distribution constant*** a statistic t such that $\mathbb{E}_\vartheta[T(Y)]$ does not depend on ϑ.

Example (Residuals of a linear model)
Residuals in a normal linear regression are first-order distribution constant but not completely independent of σ^2.

› ***Pivotal quantity*** a combinant $q(Y, \vartheta)$ whose null distribution does not depend on ϑ.

Example (Pivotal quantity)
If $Y \sim \mathcal{N}(\vartheta', 1)$, the quantity $Y_1 - \vartheta \sim \mathcal{N}(\vartheta' - \vartheta, 1)$ is a pivotal quantity when $\vartheta' = \vartheta$.

How do we solve a distribution problem? We usually employ some analytic approximations based on limit theorems or simulation methods, such as Monte Carlo methods and bootstrap methods.

Most important extensions of CLT's and LLN's are those related to martingales (**andersen1996**)

1.2 Delta method and asymptotic statistics

References **vandervaart1998** on the delta method.
severini2012 for examples of the delta method.

Theorem 19 (Delta method)
Let $\{Y_n\}$ be a sequence of random variables such that

$$\sqrt{n}(Y_n - \vartheta) \xrightarrow{d} U,$$

where U is non-degenerate. If $g(\vartheta)$ is a twice-differentiable measurable function such that $g'(\vartheta) \neq 0$, then

$$\sqrt{n}\big(g(Y_n) - g(\vartheta)\big) \xrightarrow{d} g'(\vartheta)U.$$

Extensions There are extensions of the Delta method for the case when $g'(\vartheta) = 0$, since they are simply restatements of the Lagrange form of the remainder in the Taylor series expansion.

Proof.
We can apply the mean value theorem using the two extremes

$$a = \vartheta, \quad b = \vartheta + \frac{U_n}{\sqrt{n}},$$

to write the following first-order approximation

$$g\left(\vartheta + \frac{U_n}{\sqrt{n}}\right) - g(\vartheta) = \frac{U_n}{\sqrt{n}} g'\left(\vartheta + \frac{V_n}{\sqrt{n}}\right).$$

The final step is to write

$$\sqrt{n}\big(g(Y_n) - g(\vartheta)\big) = g'(\vartheta) U_n \underbrace{\frac{g'\left(\vartheta + \frac{V_n}{\sqrt{n}}\right)}{g'(\vartheta)}}_{\xrightarrow{P} 1},$$

then we can simplify the notation by writing the second term as

$$g'(\vartheta) u_n \big(1 + o_p(1)\big),$$

where we denote $o_p(1)$ to denote a sequence such that $\xrightarrow{P} 0$. Hence, we can write (by Slutsky)

$$\begin{aligned} g'(\vartheta) U_n \big(1 + o_p(1)\big) &= g'(\vartheta) U_n + o_p(1). \\ &= g'(\vartheta)\big(U + o_p(1)\big) + o_p(1). \\ &= g'(\vartheta) U + o_p(1). \end{aligned}$$

Hence, we get the final result

$$\begin{aligned} g(Y_n) &= g(\vartheta) + \frac{1}{\sqrt{n}}\big(g'(\vartheta) U + o_p(1)\big) \\ &= g(\vartheta) + \frac{1}{\sqrt{n}} g'(\vartheta) U + \frac{1}{\sqrt{n}} o_p(1). \end{aligned}$$

□

An interesting extension of the Delta method is its multivariate generalization.

Theorem 20 (Multivariate delta method)
Id

$$\sqrt{n}(\boldsymbol{Y}_n - \vartheta) \xrightarrow{d} \mathcal{N}_d(0, \Sigma),$$

and that $g : \mathbb{R}^d \to \mathbb{R}^k$ with $k \leq d$ and twice-differentiable components $g_i(\cdot)$ for all $i = 1, \ldots, k$. Then,

$$\sqrt{n}\big(g(\boldsymbol{Y}_n) - g(\vartheta)\big) \xrightarrow{d} \mathcal{N}_k(0, D\Sigma D^\top),$$

where

$$D = (d_{ij})_{i,j} = \left(\frac{\partial g_i}{\partial \vartheta_j}\right)_{i,j}$$

Example (Exercise similar to homework)

$Y_1, \ldots, Y_n$ i.i.d $\text{Exp}(1/\vartheta)$, then $Y_i/\vartheta \sim \text{Exp}(1)$. Finding the approximate distribution of

$$\frac{\sum_{i=1}^n Y_i/\vartheta^2}{(\sum_{i=1}^n Y_i)^2/\vartheta^2} \sim \frac{\sum_{i=1}^n Z_i^2}{(\sum_{i=1}^n Z_i)^2}, \quad Z_i \sim \text{Exp}(1).$$

Using the central limit theorem we have a way of studying the following distribution,

$$\begin{pmatrix} \overline{Z}_1 \\ \overline{Z}_2 \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n Z_i/n \\ \sum_{i=1}^n Z_i^2/n \end{pmatrix},$$

since

$$\sqrt{n}(\overline{Z}_n - \mu_{\overline{Z}}) \xrightarrow{d} \mathcal{N}_2(0, \Sigma), \tag{3}$$

where we want to obtain $\mu_{\overline{Z}}$ and Σ using the expected values and variances/covariances. Afterwards, we can compute the delta method on the approximation (3).

Discussion on the bootstrap with reference to **hall1992**.

Lecture 2: Asymptotic expansions and approximations

2022-03-08

We need some tools, which are related to what we saw in the previous lectures. Mostly used for analytic approximations whose main objective are to approximating distributions, moments, and integrals.

2.1 $O_p(\cdot)$ and $o_p(\cdot)$ notation

The notation $O_p(\cdot)$ and $o_p(\cdot)$ are useful to extend to random variables the notion of order of convergence and Taylor series that we know from calculus.

Def. (o(1))
A sequence is said to be $o_p(1)$ if ...

Def. ($o(n^\alpha)$)
A sequence of random variables Y_n is said to be asymptotically of order $o(n^\alpha)$ in probability if

$$\frac{Y_n}{n^\alpha} = o_p(1).$$

Remark. If $Y_n = o_p(n^\alpha) \implies Y_n = o_p(n^{\alpha+\varepsilon})$ for any $\varepsilon > 0$.

Example (Normal)
$Y_n \sim N(0, 1/n)$ then we have that $Y_n = o_p(1)$ but it is not of order $o_p(n^{-1/2})$. To see this, we can write

$$Y_n/n^{-1/2} = \sqrt{n}Y_n \sim N(0,1) \not\xrightarrow{p} 0.$$

Example (chi square)
We have that

$$Y_n \sim \chi^2_n \implies Y_n/n^{3/2} = o_p(1).$$

Indeed, we have that $\mathbb{E}[Y_n] = n$ $\mathbb{V}[Y_n] = 2n$, therefore if ...

Def. ($O_p(1)$)
We say that Y_n is asymptotically of order $O(1)$ in probability, $O_p(1)$, if for each $\varepsilon > 0$ there exists a real number $A = A_\varepsilon > 0$ and $\bar{n} = \bar{n}_\varepsilon$ such that for every $n > \bar{n}$,

$$\mathbb{P}(|Y_n| < A) > 1 - \varepsilon.$$

Remark. An $O_p(1)$ sequence is said to be bounded in probability.

Lemma 2 ($\xrightarrow{d}$ implies boundedness)
If $Y_n \xrightarrow{d} Y$, then Y_n is bounded in probability.

Proof.
severini2012.

□

Remark. The converse is in general not true, think for example to a process that constantly switches between two bounded probability measures.

Def. ($O_p(n^\alpha)$)
A sequence is said to be asymptotically of order $O_p(n^\alpha)$ if

$$\frac{Y_n}{n^\alpha} = O_p(1).$$

Example (orders in probability)
$Y_n \sim N(0, n^{-1})$ is $O_p(n^{-1/2})$ and $Y_n \sim \chi^2_n$ is $O_p(n)$.

- If $0 < \mathbb{V}[Y_i] < \infty$ and $\mathbb{E}[Y_i] = 0$, then S_n is of order $O_p(n^{1/2})$. Example is the score function.
- If $\mathbb{E}[Y_i] \neq 0$, then $S_n = n\mu + O_p(n^{1/2})$ and so $S_n = O_p(n)$.
- If $\mathbb{V}[y_i] = \infty$, then Y_i is in the domain of attraction of a symmetric stable law and $S_n = O_p(n^{1/\vartheta})$.

Prop. 1 (Algebra of o's)
$O_p(n^a) \cdot O_p(n^b) = O_p(n^{a+b})$
$O_p(n^a) \cdot o_p(n^b) = o_p(n^{a+b})$
$o_p(n^a) \cdot o_p(n^b) = o_p(n^{a+b})$
$O_p(n^a) + O_p(n^b) = O_p(n^{\max\{a+b\}})$

Proof.
severini2012.

□

Why are moment-generating functions and characteristic functions relevant in statistics? In general, they are useful both for *specification problems* and *distribution problems*, since asymptotic results are usually easier to prove in terms of moment-generating functions and characteristic functions.

2.2 Edgeworth expansion

The Edgeworth expansion has to do with sums of random variables, and the idea is to study how we approach the limit in the central limit theorem. This can be used either to correct non-normalities and compare inferential procedures.

2.2.1 Generating functions for sums

Consider $S_n = \sum_{i=1}^n Y_i$ where the Y_i are i.i.d, then we can write the moment-generating function as

$$M_{S_n}(t) = \left(1 + \mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \dots\right)^n. \tag{4}$$

Using Newton's binomial formula,

$$(1+x)^n = \sum_{j=0}^{n} \binom{n}{j} x^j.$$

we can write (4) after truncating it at $x = \mu_1 t + \ldots + \mu_4 t^4/4!$ to get the leading terms

$$M_{S_n}(t) = 1 + n\left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!}\right) + \binom{n}{2}\left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!}\right)^2$$
$$+ \binom{n}{3}\left(\mu_1 t + \mu_2 \frac{t^2}{2!}\right)^3 + \binom{n}{4}\mu_1^4 t^4 + \dots,$$

which result in expressions for the moments of S_n in terms of the coefficients of $t^j/j!$.

It's however much easier to do this by using logarithms, and we do so by using the cumulant-generating function $K_Y(t) = \log M_Y(t)$, which can be expanded in Taylor series with the same radius of convergence as (4)

$$K_Y(t) = \kappa_1 t + \kappa_2 \frac{t^2}{2!} + \kappa_3 \frac{t^3}{3!} + \kappa_4 \frac{t^4}{4!} + \dots,$$

and the κ_r^{th} coefficient is called the cumulant of order r. Moreover, for S_n we have

$$M_{S_n}(t) = M_Y(t)^n \implies K_{S_n}(t) = nK_Y(t),$$

yielding the simple relationship

$$\kappa_r(S_n) = n\kappa_r(Y) = n\kappa_r.$$

A location change $Y \to Y + a$ affects the cumulant-generating function through a linear term, since

$$M_{Y+a}(t) = e^{at} M_Y(t) \implies K_{Y+a}(t) = at + K_Y(t),$$

hence only the first cumulant κ_1 is affected by the location change. We calculate the relationships for the cumulants $\kappa_1, \dots, \kappa_4$ by expanding $\log(1+x)$ using a Taylor series expansion,

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \dots, \quad |x| < 1, \tag{5}$$

and by applying this procedure to $K_Y(t)$ we obtain

$$K_Y(t) = \log\left(1 + \mu_1 t + \frac{\mu_2}{2!}t^2 + \frac{\mu_3}{3!}t^3 + \frac{\mu_4}{4!}t^4 + \dots\right),$$

and by using (5) we have

$$K_Y(t) = \left(\mu_1 t + \mu_2 \frac{t^2}{2!} + \mu_3 \frac{t^3}{3!} + \mu_4 \frac{t^4}{4!} + \dots\right) - \frac{1}{2}(\dots)^2 + (\dots)^3 - \frac{1}{4}\mu_1^4 t^4 + \dots, \tag{6}$$

and by rearranging the terms and discarding those of order 5 or more we obtain the cumulants of the cumulant-generating function using

$$\kappa_1 = \mu$$

$$\kappa_2 = \mu_2 - \mu_1^2$$

$$\kappa_1 = \mu_3 - 3\mu_1\mu_2 + 2\mu_1^3 = \mathbb{E}[Y - \mu_1]^3$$

$$\kappa_1 = \mathbb{E}[Y - \mu_1]^4 - 3\mathbb{V}[Y]^2$$

Change of variables. When applying a scale change,

$$K_{Y/b}(t) = K_Y(t/b) \implies \kappa_r' = \kappa_r / b^r.$$

Since the cumulants are affected by a scale change, we can standardize them using the standardized cumulants

$$\rho_r = \kappa_r / \kappa_2^{r/2} = \kappa_r / \sigma^r, \quad r = 3, 4, \dots,$$

which are invariant under affine transformations $a + bY$, i.e.

$$\rho_r = \kappa_r \left(\frac{Y - \mu}{\sigma}\right).$$

There are constraints between cumulants given by some inequalities such as $\rho_4 \geq \rho_3^2 - 2$.

Example (Multivariate normal distribution)

2.2.2 Generating functions for a standardized sum

Let $S_n^* = (S_n - n\mu)/\sqrt{n\sigma^2}$ be the standardized sum of n i.i.d random variables, then if $\kappa_1 = \mathbb{E}[Y]$ and $\kappa_2 = \sigma^2 = \mathbb{V}[Y]$, we have that

$$\kappa_r(S_n^*) = \kappa_r\left(\frac{S_n}{\sqrt{n\sigma^2}}\right) = \frac{\kappa_r(S_n)}{n^{r/2}\sigma^r} = \frac{n\kappa_r}{n^{r/2}\sigma^r} = \frac{\rho_r}{n^{r/2-1}}, \quad r = 3, 4, \dots,$$

where κ_r and ρ_r are the cumulants and the standardized cumulants of Y. In particular, under random sampling,

$$\rho_1(S_n^*) = 0$$

$$\rho_2(S_n^*) = 1$$

$$\rho_r(S_n^*) = O\big(n^{-r/2+1}\big), \quad r = 3, 4, 5, \dots$$

and as $n \to \infty$ we have that the cumulants of S_n^* approach the cumulants of $N(0, 1)$.

Therefore, as n diverges the expansion of (6) for S_n^* can be written as

$$\begin{aligned} K_{S_n^*}(t) &= \overbrace{\kappa_1(S_n^*)}^{=0} t + \overbrace{\frac{\kappa_2(S_n^*)}{2!}}^{=1} t^2 + \frac{\kappa_3(S_n^*)}{3!}t^3 + \frac{\kappa_4(S_n^*)}{4!}t^4 + O(n^{-3/2}) \\ &= \frac{1}{2}t^2 + \frac{\rho_3}{6\sqrt{n}}t^3 + \frac{\rho_4}{24n}t^4 + \underbrace{O(n^{-3/2})}_{\kappa_5=\rho_5/n^{3/2}}. \end{aligned} \tag{7}$$

and by using $e^x \approx 1 + x + \frac{x^2}{2}$, we can write the moment-generating function fun (7),

$$\begin{aligned} M_{S_n^*}(t) &= e^{t^2/n} \cdot e^{\frac{\rho_3}{6\sqrt{n}}t^3 + \frac{\rho_4}{24}t^4 + O(n^{-3/2})} \\ &= e^{t^2/n}\left\{1 + \frac{\rho_3}{6\sqrt{n}}t^3 + \frac{\rho_4}{24n}t^4 + \frac{\rho_3^2}{72n}t^6 + O(n^{-3/2})\right\}. \end{aligned} \tag{8}$$

We can see from (8) that the approximation is the moment-generating function of a standard normal distribution plus some corrections due to skewness and kurtosis. Indeed, if a distribution is symmetric then $\rho_3 = 0$ and convergence to the moment-generating function of a normal distribution is of order $O(n^{-1})$ instead of $O(n^{-1/2})$.

Def. (Edgeworth expansion)

We define the Edgeworth expansion of Y by inverting equation (8) to recover the density that generated it

$$M_{S_n^*}(t) = \int_{-\infty}^{\infty} e^{ty}\left\{p_{S_n^*}^E(y) + O(n^{-3/2})\right\} dy, \tag{9}$$

so that we get

$$p_{S_n^*}(y) = p_{S_n^*}^E(y) + O(n^{-3/2}). \tag{10}$$

Remark. This can be done explicitly by introducing Hermite polynomials,

$$H_r(y)\varphi(y) = (-1)^r \frac{\partial^r \varphi(y)}{\partial y^r},$$

where $\varphi(y)$ is the probability density function of a $N(0,1)$ random variable. The first term is obtained by using $\frac{\partial \varphi(y)}{\partial y} = -y\varphi(y)$.

Remark. Even Hermite polynomials are even functions, odd Hermite polynomials are odd functions.

For $r \in \mathbb{N}$, we have that

$$\int_{-\infty}^{\infty} e^{ty} H_r(y)\varphi(y)dy = t^r e^{t^2/2} \tag{P1}$$

and

$$\int_{-\infty}^{y} \varphi(t)H_r(t)\,\mathrm{d}t = -\varphi(y)H_{r-1}(y) \tag{P2}$$

Example (Hermite polynomial)
The term $e^{\frac{1}{2}t^2}\frac{\rho_3}{6\sqrt{n}}t^3$ in (8) can be rewritten by applying (P1) with $r = 3$,

$$e^{\frac{1}{2}t^2}\frac{\rho_3}{6\sqrt{n}}t^3 = \int_{-\infty}^{\infty} e^{ty}\frac{\rho_3}{6\sqrt{n}}H_3(y)\varphi(y)dy.$$

Applying the property to all terms in (8) yields the following approximation for the density,

$$p^E_{S_n^*} = \varphi(y)\left\{1 + \frac{\rho_3}{6\sqrt{n}}H_3(y) + \frac{\rho_4}{24n}H_4(y) + \frac{\rho_3^2}{72n}H_6(y) + O(n^{-3/2})\right\}. \tag{11}$$

Remark. The above representation is more precise at $y = 0$ because $H_3(y)$ also vanishes.

Using (P2) we also obtain an Edgeworth expansion for F,

$$F^E_{S_n^*}(y) = \Phi(y) - \varphi(y)\left\{\frac{\rho_3}{6\sqrt{n}}H_2(y) + \frac{\rho_4}{24n}H_3(y) + \frac{\rho_3^2}{72n}H_5(y)\right\} + O(n^{-3/2}). \tag{12}$$

Validity. For Y with finite moments up to the 5-th order, conditions for validity of (10) are

- *Continuous case*: if there exists a value $\bar{n} \in \mathbb{N}^+$ such that the density $p_{S_n^*}(y)$ is continuous and bounded for every $n > \bar{n}$;
- *Discrete case*: more complicated.

Accuracy. Uniform bound in the absolute error of order $O(n^{-3/2})$, but no bound is available in terms of relative error. Hence, the Edgeworth approximation yields bad results in the tails of the distribution, hence it is used as a tool to prove formal results or to obtain better results.

2.2.3 Cornish-Fisher expansion

We can invert the approximation $F^E_{S_n^*(y)}$ to get an approximation for the α^{th} quantile of S_n^* in terms of the quantiles u_α of $N(0,1)$. The resulting expansion is called the Cornish-Fisher expansion,

$$y_\alpha = u_\alpha + \frac{\rho_3}{6\sqrt{n}}(u_\alpha^2 - 1) + \frac{\rho_4}{24n}(u_\alpha^3 - 3u_\alpha) - \frac{\rho_3^2}{36n}(2u_\alpha^3 - 5u_\alpha) + O(n^{-\frac{3}{2}}), \tag{13}$$

and its inverse is

$$u_\alpha = y_\alpha - \frac{\rho_3}{6\sqrt{n}}(y_\alpha^2 - 1) - \frac{\rho_4}{24n}(y_\alpha^3 - 3y_\alpha) + \frac{\rho_3^2}{36n}(4y_\alpha^3 - 7y_\alpha) + O(n^{-3/2}). \tag{14}$$

Remark. Note that if $\alpha \sim U(0,1)$ and u_α is a realization of $U \sim N(0,1)$, then y_α is a realization of $Y = S_n^*$. Hence, the inverse Cornish-Fisher expansion (14) can be read as

$$U(Y) = Y - \frac{\rho_3}{6\sqrt{n}}(Y^2 - 1) + \frac{\rho_4}{24n}(Y^3 - 3Y) + \frac{\rho_3^3}{36n}(4Y^3 - 7Y) + O_p(n^{-3/2}), \tag{15}$$

where $O(n^{-3/2}) \to O_p(n^{-3/2})$ since they are polynomials in Y. Hence, (15) is a polynomial normalizing transformation for $Y = S_n^*$.
The Edgeworth approximation permits for instance

1. To assess the coverage accuracy of confidence limits based on approximate pivotal quantities.
2. To assess the coverage accuracy of bootstrap confidence limits.

Example (Use of the Edgeworth approximation)
Consider an approximate pivotal quantity

$$Q_n = (\widehat{\vartheta}_n - \vartheta)/\sqrt{v} \dot{\sim} N(0,1),$$

and that an Edgeworth expansion (12) holds for the distribution function of Q_n under ϑ,

$$\mathbb{P}_\vartheta(Q_n \leq q) = \Phi(q) + \frac{1}{\sqrt{n}} h(q) + O(n^{-1}),$$

where $h(\cdot)$ is an even function. Now, the upper confidence limit for ϑ is

$$\widehat{\vartheta} + u_{1-\alpha}\sqrt{v},$$

which has coverage $1 - \alpha + O(n^{-1/2})$. Indeed,

$$\begin{aligned}
\mathbb{P}_\vartheta(\vartheta \leq \widehat{\vartheta}_n + u_{1-\alpha}\sqrt{v}) &= \mathbb{P}_\vartheta(Q_n \geq -u_{u_1-\alpha}) \\
&= 1 - P_\vartheta(\ldots) \\
&= 1 - \Phi(-u_{1-\alpha}) + O(n^{-1/2}) \\
&= 1 - \alpha + O(n^{-1/2})
\end{aligned}$$

On the other hand, the confidence interval $\widehat{\vartheta} \pm u_{1-\frac{\alpha}{2}}\sqrt{v}$ has coverage

$$\begin{aligned}
\mathbb{P}_\vartheta(-u_{1-\alpha/2} \leq \frac{\widehat{\vartheta}_n - \vartheta}{\sqrt{v}} \leq u_{1-\alpha/2}) &= \Phi(u_{1-\alpha/2}) + \frac{h(u_{1-\frac{\alpha}{2}})}{\sqrt{n}} - \Phi(-u_{1-\frac{\alpha}{2}}) - \frac{h(-u_{1-\frac{\alpha}{2}})}{\sqrt{n}} + O(n^{-1}) \\
&= 1 - \alpha + 0 + O(n^{-1}),
\end{aligned}$$

since h is an even function of the arguments.

Example (hall1992)
It's much better to first standardize the statistic (*pre-pivoting*) before applying the bootstrap procedure, since this improves the order of approximation from $O(n^{1/2})$ to $O(n^{-1})$.

Assuming consistency (**davison1997**), i.e.

$$\mathbb{P}_{\widehat{F}_n}(Q^* \leq q) \longrightarrow \mathbb{P}_F(Q_n \leq q),$$

the Edgeworth expansion yields

$$\mathbb{P}_F(Q_n \leq q) = \Phi(q) + \frac{h(q)}{\sqrt{n}} + O(n^{-1}),$$

the corresponding expansion for Q^* is

$$\mathbb{P}_{\widehat{F}_n}(Q_n \leq q) = \Phi(q) + \frac{\widehat{h}(q)}{\sqrt{n}} + O_p(n^{-1}),$$

and typically $\widehat{h}(q) = h(q) + O_p(n^{-1/2})$ so that

$$\frac{1}{\sqrt{n}}\widehat{h}(q) = \frac{1}{\sqrt{n}}(h(q) + O_p(n^{-1/2})) = \frac{1}{\sqrt{n}}h(q) + O_p(n^{-1}).$$

$$\mathbb{P}_{\widehat{F}_n}(Q^* \leq q) \dots$$

Had we not studentized, the error would have been $O_p(n^{-1/2})$ (slides).

2.3 Approximations of moments

After discussing approximations to the distribution, we turn to transforming the pivotal statistic to improve the convergence $O(n^{-1/2})$ to the normal distribution. Since asymptotic normality is preserved under smooth transformations, we try to improve the normal approximation though a transformation which, for instance, might yield $\rho_3 = 0$ and thus improve the convergence in (10).

Def. (variance-stabilizing parametrization)
Let T_n be a scalar statistic such that $T_n \dot{\sim} N(\vartheta, \sigma^2(\vartheta)/n)$, then $\psi = g(\vartheta)$ is called a variance-stabilizing parametrization if

$$\sigma(\vartheta)g'(\vartheta) = c \iff g(\vartheta) = \int_{\vartheta_0}^{\vartheta} \sigma(t)^{-1} dt,$$

where c is a constant. Indeed,

$$g(T_n) \dot{\sim} N(g(\vartheta), c).$$

Remark. The main objectives of variance stabilization are

- making the (asymptotic) accuracy of an estimator independent of the parameter ϑ;
- approach a situation of equal response variance if T_n plays the role of response variable in a regression model.

Example (Fisher's z transformation)
If r is the sample correlation coefficient, then

$$z = \frac{1}{2}\log\frac{1+r}{1-r}$$

has a $\rho_3(z)$ if order $O(n^{-2})$ and the error in the normal approximation for Z is of order $O(n^{-1})$,

$$z \dot{\sim} N\left(\frac{1}{2}\log\frac{1+\rho}{1-\rho}, \frac{n-1}{n-3}\right),$$

and therefore it's possible to derive $\mathbb{P}(r \leq r_0; \rho)$ by inverting the Z transform with a high degree of accuracy.

Model	T_n	$\frac{\sigma^2(\theta)}{n}$	$\psi(\theta)$	$\frac{\sigma^2(\theta)}{n}\left(\psi'(\theta)\right)^2$	Author
Y_i, i.i.d., $i = 1, \ldots, n$, $P(\theta)$	$\bar{Y}_n$	$\frac{\theta}{n}$	$\sqrt{\theta}$	$\frac{1}{4n}$	Bartlett (1936b) Anscombe (1948)
Y_i, i.i.d., $i = 1, \ldots, n$, $Bi(1, \theta)$	$\bar{Y}_n$	$\frac{\theta(1-\theta)}{n}$	$\arcsin\sqrt{\theta}$	$\frac{1}{4n}$	Anscombe (1948)
Y_i, i.i.d., $i = 1, \ldots, n$, $Ga(1, \theta)$	$\bar{Y}_n$	$\frac{\theta}{n}$	$\sqrt{\theta}$	$\frac{1}{4n}$	Fisher (1922b)
(X_i, Y_i) i.i.d., $i = 1, \ldots, n$, bivariate normal with correlation coefficient ρ	r	$\frac{(1-\rho^2)^2}{n}$	$\text{arctgh}\rho = \frac{1}{2}\log\frac{1+\rho}{1-\rho}$	$\frac{1}{n}$	Fisher (1921)

Figure 1: Examples of variance stabilizing transformations for some common distributions.

Remark. In the multivariate case, it is not possible to make the variance-covariance matrix proportional to the identity matrix, and it has been proven by **holland1973**.

Using the Taylor formula we can write

$$g(Y_n) = g(\vartheta)\frac{1}{\sqrt{n}}g'(\vartheta)U + o_p\left(\frac{1}{\sqrt{n}}\right),$$

and we can generalize it to higher orders

Def. (stochastic Taylor formula)
Suppose $\{Y_n\}$ is a sequence of random variables such that

$$Y_n = c + O_p(n^{-\alpha}), \quad \alpha > 0,$$

and $f : \mathbb{R} \longrightarrow \mathbb{R}$ has continuous partial derivatives up to order $k+1$, then

$$f(Y_n) = f(c) + \sum_{m=1}^{k}\frac{1}{m!}\frac{\partial^m}{\partial y^m}f(y)\Big|_{y=c}(Y_n - c)^m + O_p\big(n^{-(k+1)\alpha}\big), \tag{16}$$

provided that $f(\cdot)$ does not depend on n, since $f^{(m)}(y) = O(1)$. Equation (**??**) is called the stochastic Taylor formula .

Example (approximation for $\mathbb{E}(1/\overline{Y}_n)$)
We have that

$$\mathbb{E}\left[\frac{1}{\overline{Y}_n}\right] = \mathbb{E}\left[\frac{1}{\mu} + \frac{1}{\mu^2}(\overline{Y}_n - \mu) + \frac{1}{2}\cdot\frac{2}{\mu^3}(\overline{Y}_n - \mu)^2 + O_p\big((\overline{Y}_n - \mu)^3\big)\right].$$

Def. (multivariate Taylor formula)
If $Y_n = c + O_p(n^{-\alpha})$ and $f : \mathbb{R}^d \longrightarrow \mathbb{R}$, is a smooth function with continuous derivatives up to order $k+1$, then the multivariate Taylor formula yields

$$f(Y_n) = f(x) + \sum_{m=1}^{k} \frac{1}{m!} f_{I_m}(c)(Y_n - c)_{I_m} + O_p\Big(n^{-(k+1)\alpha}\Big),$$

where $(Y_n - c)_{I_m} = (Y_n - c)_{i_1} \cdots (Y_n - c)_{i_m}$ and

$$f_{I_m}(y) = \frac{\partial^m}{\partial y_{i_1} \cdots \partial y_{i_m}} f(y)$$

Convention. Einstein's summation convention: when an index appears two or more times in a product of elements, then the summation over the range of that index is understood:

$$(Y_n - c)_{i_1} f_{i_1}(x) = \sum_{i_1=1}^{d} (Y_n)(Y_n - c)_{i_1} f_{i_1}(c).$$

If $Y_n = \overline{Y}_n$ and $c = \mu$ it can be shown that

- $\mathbb{E}[(\overline{Y}_n - \mu)_{I_m}] = O(n^{-m/2})$ if m is even
- $\mathbb{E}[(\overline{Y}_n - \mu)_{I_m}] = O(n^{-(m+1)/2})$ if m is even

Hence,

$$\kappa_3(\overline{Y}_n) = O(n^{-2})$$
$$\kappa_4(\overline{Y}_n) = O(n^{-2})$$
$$\kappa_5(\overline{Y}_n) = O(n^{-5/2})$$
$$\kappa_6(\overline{Y}_n) = O(n^{-5/2})$$

Theorem 21
Let $f(\cdot)$ be a smooth function with $f'(\mu) \neq 0$ and $f''(\mu) \neq 0$, then

$$\mathbb{E}[(f(\overline{Y}_n) - \mathbb{E}[f(\overline{Y}_n)])^3] = \frac{\bar{\mu}_3(f'(\mu)^3) + 3\sigma^4 f'(\mu)^2 f''(\mu)}{n^2} + O(n^{-3}). \tag{17}$$

Proof.
pace1997

□

Remark. Finding a solution to the differential equation in (**??**) is hard unless we restrict for example to functions of the form

$$f(y) = y^h,$$

which yield some useful transformation.

Example (Anscombe)

Remark. Restricting to $f(y) = ay^2 + by + c$ we obtain an explicit solution, and by defining the standardized variable Z_n then

$$f(\overline{Y}_n) = z_n - \dots$$

Example (quadratic normalizing transformation)

2.4 Laplace approximation

We want to approximate an integral of the form

$$I(n) = \int_{\mathbb{R}} \exp\{-ng(y)\}\, dy,$$

where $g(\cdot)$ is a smooth real function with a minimum at $\tilde{y}$ so that $g'(\tilde{y}) = 0$ and $g''(\tilde{y}) > 0$. In general, this is particularly useful in Bayesian inference for integrals of the form

$$\int e^{\ell(\vartheta)} \pi(\vartheta) d\vartheta,$$

where $\ell(\vartheta) = -n\left(-\frac{1}{n}\sum_{i=1}^n \log p(y_i|\vartheta)\right) = O_p(n)$, and the $\tilde{y}$ will be the mode of the posterior distribution of $\vartheta|Y$.

The asymptotic behaviour of $I(n)$ is entirely determined by the local behaviour of $g(\cdot)$ in a neighbourhood of $\tilde{y}$,

To obtain the result we can apply a Taylor approximation of $g(y)$ around $\tilde{y}$,