TECHNISCHE UNIVERSITÄT WIEN

Introduction to Statistics
The Univariate and Multivariate Delta Method
Sampling Theory

LV Nr. 105.692
Summer Semester 2021

The Delta Method

Motivating Example: Estimating the odds

- Assume we observe $X_1, \ldots, X_n$ independent Bernoulli(p) random variables.
- Typically, we are interested in the parameter p but there is also interest in the *odds*, i.e.

$$\frac{p}{1-p}$$

- For example, if the outcomes of a medical treatment occur with $p = \frac{2}{3}$, then the odds of getting better is $2:1$.
- If there is another treatment with success probability r, we might also be interested in the relative odds of one treatment over another, i.e. the ratio

$$\frac{p}{1-p} \Big/ \frac{r}{1-r}$$

Estimating the odds

- If we wished to estimate p, we would typically estimate this quantity with the observed success probability

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

- To estimate the odds, it then seems perfectly natural to use

$$\frac{\hat{p}}{1 - \hat{p}}$$

  as an estimate for $\frac{p}{1-p}$.
- Although we know how to compute the variance of the estimator $\hat{p} = \bar{X}$, the question is what the variance of $\frac{\hat{p}}{1-\hat{p}} = \frac{\bar{X}}{1-\bar{X}}$ is, or how can we approximate its sampling distribution?
- The Delta method gives a technique for doing this.

The Delta Method

- The name *delta* refers to differentiation as in $\Delta y/\Delta x$
- It is a joint application of Slutsky's theorem, the continuous mapping theorem and Taylor series expansion

Theorem

Suppose

$$n^{\alpha}(X_n - \theta) \xrightarrow{d} Y$$

where $\alpha > 0$. *Let* g *be a function differentiable at* θ, *then*

$$n^{\alpha}(g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Y$$

Proof

Why? The assumption that g is differentiable at θ means

$$g(\theta + h) = g(\theta) + g'(\theta)h + o(h)$$

where here the "little oh" of h refers to $h \to 0$ rather than $h \to \infty$ (h is a function of the sample size that goes to ∞). We can write it as

$$o(h) = |h|\psi(h) \quad \text{where } \psi(h) \to 0 \quad \text{as } h \to 0$$

This implies

$$n^{\alpha}(g(X_n) - g(\theta)) = g'(\theta)n^{\alpha}(X_n - \theta) + n^{\alpha}o(X_n - \theta)$$

and the first term on the right-hand side converges to $g'(\theta)Y$ by the continuous mapping theorem.

Proof (contd.)

We can rewrite the second term on the right-hand side

$$|n^{\alpha}(X_n - \theta)|\psi(X_n - \theta)$$

By the continuous mapping theorem,

$$|n^{\alpha}(X_n - \theta)| \xrightarrow{d} |Y|$$

and

$$X_n - \theta \xrightarrow{P} 0$$

by Slutsky's theorem since $n^{\alpha}(X_n - \theta) \xrightarrow{d} Y$. Hence,

$$\psi(X_n - \theta) \xrightarrow{P} 0$$

by the continuous mapping theorem.

Proof (contd.)

Putting this all together,

$$|n^{\alpha}(X_n - \theta)|\psi(X_n - \theta) \xrightarrow{P} 0$$

by Slutsky's theorem.
Another application of Slutsky's gives

$$n^{\alpha}(g(X_n) - g(\theta)) = g'(\theta)n^{\alpha}(X_n - \theta) + n^{\alpha}o(X_n - \theta) \xrightarrow{d} g'(\theta)Y$$

The Delta Method (contd.)

Let $X_1, \ldots, X_n$ be iid $\text{Exp}(\lambda)$ random variables with expectation $1/\lambda$.
Let

$$\overline{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Then the CLT obtains

$$\sqrt{n} \left(\overline{X}_n - \frac{1}{\lambda} \right) \xrightarrow{d} \mathcal{N} \left(0, \frac{1}{\lambda^2} \right)$$

Qn: What is the asymptotic distribution of $1/\overline{X}_n$?

In this case,

$$g(x) = \frac{1}{x}$$
$$g'(x) = -\frac{1}{x^2}$$

The Delta Method (contd.)

$$\sqrt{n}\left(\overline{X}_n - \frac{1}{\lambda}\right) \xrightarrow{d} Y$$

$$\begin{aligned}\sqrt{n}\left(g(\overline{X}_n) - g(\frac{1}{\lambda})\right) &= \sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \\ &\xrightarrow{d} g'\left(\frac{1}{\lambda}\right) Y \\ &= -\lambda^2 Y\end{aligned}$$

Recall that the random variable Y had the $\mathcal{N}(0, 1/\lambda^2)$ distribution. Since a linear function of a normal is normal,

$$-\lambda^2 Y \sim \mathcal{N}\left(-\lambda^2 \mathbb{E}(Y), (-\lambda^2)^2 \mathbb{V}\mathrm{ar}(Y)\right) = \mathcal{N}(0, \lambda^2)$$

Hence we have finally arrived at

$$\sqrt{n}\left(\frac{1}{\overline{X}_n} - \lambda\right) \xrightarrow{d} \mathcal{N}(0, \lambda^2)$$

The Delta Method (contd.)

Since we routinely use the delta method in the case where the rate is $\sqrt{n}$ and the limiting distribution is normal, it is worthwhile working out some details of that case.

Lemma

Let X_n be a sequence of random variables such that

$$\sqrt{n}(Y_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Suppose g is a function differentiable at θ, i.e., $g'(\theta)$ exists and $g'(\theta) \neq 0$. Then,

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} \mathcal{N}(0, \sigma^2 (g'(\theta))^2)$$

Estimating the odds

By the CLT,

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

where p is the binomial success probability.
We take $g(p) = \frac{p}{1-p}$ so that $g'(p) = \frac{1}{(1-p)^2}$ and

$$\begin{aligned}\mathbb{V}\mathrm{ar}\left(\frac{\hat{p}}{1-\hat{p}}\right) &\approx g'(p)^2 \; \mathbb{V}\mathrm{ar}\left(\hat{p}\right) \\ &= \left(\frac{1}{(1-p)^2}\right)^2 \cdot \frac{p(1-p)}{n} = \frac{p}{n(1-p)^3},\end{aligned}$$

giving us an approximation for the variance of our estimator. Thus,

$$\sqrt{n}\left(\frac{\hat{p}}{1-\hat{p}} - \frac{p}{1-p}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p}{(1-p)^3}\right)$$

“Sloppy” Delta method

We can turn this into a “sloppy” version of the delta method. If

$$X_n \approx \mathcal{N}\left(\theta, \frac{\sigma^2}{n}\right)$$

then

$$g(X_n) \approx \mathcal{N}\left(g(\theta), \frac{(g'(\theta))^2\sigma^2}{n}\right)$$

In particular, if we start with the “sloppy version” of the CLT

$$\overline{X}_n \approx \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

we obtain a “sloppy” version of the delta method

$$g(\overline{X}_n) \approx \mathcal{N}\left(g(\mu), \frac{(g'(\mu))^2\sigma^2}{n}\right)$$

The Delta Method (contd.)

- Be careful not to think of the last special case as all there is to the delta method, since the delta method is really much more general.
- The delta method turns one convergence in distribution result into another.
- The first convergence in distribution result need not be the CLT. The parameter in the general theorem need not be the mean.

Application: Approximate expectation and variance

1. Let X be a random variable with $\mathbb{E}(X) = \mu \neq 0$. We want to approximate a function $g(\mu)$ by a first-order approximation

$$g(x) \approx g(\mu) + g'(\mu)\,(X - \mu)$$

Assume $g'(\mu) \neq 0$. Then, using $g(X)$ *as an estimator of* $g(\mu)$,

$$\mathbb{E}\, g(X) \approx g(\mu)$$
$$\mathbb{V}\mathrm{ar}\, g(X) \approx (g'(\mu))^2\, \mathbb{V}\mathrm{ar}\, X.$$

2. Particularly, let us take $g(\mu) = \frac{1}{\mu}$, with μ unknown. If we estimate $\frac{1}{\mu}$ with $\frac{1}{X}$, then

$$\mathbb{E}\left(\frac{1}{X}\right) \approx \frac{1}{\mu}$$
$$\mathbb{V}\mathrm{ar}\left(\frac{1}{X}\right) \approx \frac{1}{\mu^4}\, \mathbb{V}\mathrm{ar}\, X.$$

Approximate expectation and variance

Consider now the (sample) mean $\bar{X}$ of a random iid (sample) $X_1, \ldots, X_n$. For $\mu \neq 0$ we have

$$\sqrt{n}\left(\frac{1}{\bar{X}} - \frac{1}{\mu}\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{1}{\mu^4}\,\mathbb{V}\mathrm{ar}(X_1)\right).$$

1. If we do not know $\mathbb{V}\mathrm{ar}(X_1)$, in order to use the Delta method and the approximation above, we need to estimate the variance. For that, we can use *sample* variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

2. Additionally, if we also do not know μ, we have to estimate $\frac{1}{\mu}$ as well. Thus, by estimating all unknown terms, we obtain the approximate variance

$$\widehat{\mathbb{V}\mathrm{ar}}\left(\frac{1}{\bar{X}}\right) \approx \left(\frac{1}{\bar{X}}\right)^4 \cdot S^2.$$

Moreover, since both $\bar{X}$ and S^2 are consistent estimators, i.e. $\bar{X} \to \mu$ and $S^2 \to \sigma^2$ in probability, after applying Slutsky's theorem we conclude

$$\frac{\frac{1}{\bar{X}} - \frac{1}{\mu}}{\frac{1}{\sqrt{n}} \cdot \left(\frac{1}{\bar{X}}\right)^2 \cdot S} \xrightarrow{\mathrm{d}} \mathcal{N}(0,1).$$

Let $X_1, X_2, \ldots$ be a sequence of i.i.d. random variables with $\mathbb{E}(X_1) = \mu$ and $\mathbb{V}\mathrm{ar}(X_1) = \sigma^2$.
What does the Delta method tell us about the asymptotic distribution of $\bar{X}_n^2$?
Since $g(x) = x^2$ and $g'(x) = 2x$ we have by the delta method

$$\frac{\bar{X}_n^2 - \mu^2}{\frac{2\mu\sigma}{\sqrt{n}}} \xrightarrow{d} \mathcal{N}(0,1).$$

Therefore, for large n, $\bar{X}_n^2$ is approximately normal with mean μ^2 and variance $\frac{4\mu^2\sigma^2}{n}$.

A generalization of Lemma 2 for the case $g'(\mu) = 0$.

If $g'(\mu) = 0$ then we use one more term in the Taylor expansion to obtain

$$\begin{aligned} g(X_n) &= g(\theta) + g'(\theta)\,(X_n - \theta) + \frac{g''(\theta)}{2}\,(X_n - \theta)^2 + R_2 \\ &= g(\theta) + \frac{g''(\theta)}{2}\,(X_n - \theta)^2 + R_2 \end{aligned}$$

where the remainder $R_2 \to 0$ as $Y_n \to \theta$.
We recall that the square of a standard normal distribution is a χ_1^2 (chi squared random variable with one degree of freedom).
This implies

$$\frac{(X_n - \theta)^2}{\frac{\sigma^2}{n}} \xrightarrow{d} \chi_1^2.$$

Second-order Delta method

Lemma

Let X_n be a sequence of random variables such that

$$\sqrt{n}(X_n - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Suppose for a function g and a specific value of θ, $g'(\theta) = 0$ and $g''(\theta)$ exists and $g''(\theta) \neq 0$. Then,

$$\sqrt{n}(g(Y_n) - g(\theta)) \xrightarrow{d} \sigma^2 \frac{g''(\theta)}{2} \chi_1^2$$

Back to Example 18. If $\mu = 0$ then the normal limit in (16) is degenerate, i.e. the expression (16) merely states that $\sqrt{n}\,\bar{X}_n^2$ converges in probability to a constant 0, which is not what is meant by the *asymptotic distribution*. Thus, the case $\mu = 0$ is treated separately, i.e. the previous theorem should be used. Thus,

$$n\,\bar{X}_n^2 \xrightarrow{\text{d}} \sigma^2 \chi_1^2.$$

In the following example, Lemma 3 should be applied in order to estimate the variance of a binomial random variable.

Example

[HW] Let $X_1, X_2, \ldots$ be independent Bernoulli(p) random variables and let $Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

(a) Show that $\frac{Y_n - p}{\frac{1}{\sqrt{n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, p(1-p))$.

(b) Show that for $p \neq 0.5$ the estimate of variance $Y_n(1-Y_n)$ satisfies

$$\frac{Y_n(1-Y_n) - p(1-p)}{\frac{1}{\sqrt{n}}} \xrightarrow{\mathrm{d}} \mathcal{N}(0, (1-2p)^2 p(1-p)).$$

(c) Show that for $p = 0.5$ it holds $\frac{Y_n(1-Y_n) - \frac{1}{4}}{\frac{1}{n}} \xrightarrow{\mathrm{d}} -\frac{1}{4}\chi_1^2$.

Variance Stabilizing Transformations

An important application of the delta method is variance stabilizing transformations. The idea is to find a function g such that the limit in the delta method

$$n^{\alpha}g(X_n) - g(\theta)) \xrightarrow{d} g'(\theta)Y$$

has variance that does not depend on the parameter θ.
The asymptotic variance is

$$\mathbb{V}\text{ar}(g'(\theta)Y) = (g'(\theta))^2\mathbb{V}\text{ar}(Y)$$

so for this problem to make sense $\mathbb{V}\text{ar}(Y)$ must be a function of θ and no other parameters.
Thus variance stabilizing transformations usually apply only to a distributions with a single parameter.

Variance Stabilizing Transformations

Write

$$\mathbb{V}\mathrm{ar}_\theta(Y) = v(\theta)$$

Then, we try to find g such that

$$(g'(\theta))^2 v(\theta) = c$$

or, equivalently,

$$g'(\theta) = \frac{c}{v(\theta)^{1/2}}$$

The fundamental theorem of calculus ascertains that any indefinite integral of the right-hand side will do.

Example: Poisson

Let $X \sim \mathcal{P}(\lambda)$. Then by the CLT the random variable $\frac{X-\lambda}{\sqrt{\lambda}}$ is asymptotically normal $\mathcal{N}(0,1)$. Find a transformation g such that $g(X)$ is asymptotically $\mathcal{N}(g(\lambda), c^2)$ where $c > 0$ is a suitable constant. The asymptotic is for $\lambda \to \infty$.

Solution: By the Delta method,

$$g(X) \approx \mathcal{N}(g(\lambda), (g'(\lambda)^2\,\lambda)).$$

Setting $g'(\lambda)^2\,\lambda = c^2$ we obtain $g'(\lambda) = \frac{c}{\sqrt{\lambda}}$. Thus,

$$g(x) = c \cdot \int \frac{dx}{\sqrt{x}} = 2c\,\sqrt{x}.$$

If we choose $c = \frac{1}{2}$, then the transformation is $g(x) = \sqrt{x}$ and we have

$$\sqrt{X} \approx \mathcal{N}(\sqrt{\lambda}, \frac{1}{4}).$$

Example: Bernoulli

Let X_i be iid $\mathcal{B}ernoulli(p)$. Then by the CLT,

$$\sqrt{n}(\overline{X}_n - p) \xrightarrow{d} \mathcal{N}(0, p(1-p))$$

We need to find an indefinite integral of $c/\sqrt{p(1-p)}$.
Let $p = (1+w)/2$. Then,

$$\int \frac{cdp}{\sqrt{p(1-p)}} = \int \frac{cdw}{\sqrt{1-w^2}} = c \text{ arcsine}(w) + d$$

where d is an arbitrary constant and arcsine is the inverse of sine.
Thus,

$$g(p) = \text{arcsine}(2p-1), \quad 0 \leqslant p \leqslant 1$$

is a variance stabilizing transformation for the Bernoulli distribution. We check this:

$$g'(p) = \frac{1}{\sqrt{p(1-p)}}$$

so that

$$\sqrt{n}(g(\overline{X}_n) - g(p)) \xrightarrow{d} \mathcal{N}(0, 1)$$

Variance Stabilizing Transformations (cont.)

It is important that the parameter θ in the discussion of variance stabilizing transformations is as it appears in the convergence distribution result we start with

$$n^{\alpha}(X_n - \theta) \xrightarrow{d} Y$$

Specifically, if we start with the CLT,

$$\sqrt{n}(\overline{X}_n) - \mu) \xrightarrow{d} Y$$

θ must be the mean μ.
We need to find an indefinite integral of $v(\mu)^{-1/2}$, where $v(\mu)$ is the variance expressed as a function of the mean, not some other parameter.

Example: Geometric Assume X_i are iid Geometric(p) with $\mathbb{E}(X_i) = (1-p)/p$, $\mathbb{V}\text{ar}(X_i) = (1-p)/p^2$. Find the variance stabilizing transformation of $\overline{X}_n$.

Multivariate Convergence in Probability

We define the length of a vector $\mathbf{x} = (x_1, \ldots, x_k)^T$ as

$$||\mathbf{x}|| = \sqrt{\mathbf{x}^T\mathbf{x}} = \sqrt{\sum_{i=1}^{k} x_i^2}$$

Then we say a sequence $\mathbf{X}_1, \mathbf{X}_2, \ldots$ of random vectors converges in probability to a constant vector $\mathbf{a} = (a_1, \ldots, a_k)^T$ if

$$||\mathbf{X}_n - \mathbf{a}|| \xrightarrow{P} 0$$

which by the continuous mapping theorem happens if and only if

$$||\mathbf{X}_n - \mathbf{a}||^2 = \sum_{i=1}^{k} (X_{ni} - a_i)^2 \xrightarrow{P} 0$$

Multivariate Convergence in Probability (cont.)

We write

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}$$

or,

$$\mathbf{X}_n - \mathbf{a} = o_p(1)$$

to denote

$$||\mathbf{X}_n - \mathbf{a}||^2 \xrightarrow{P} 0$$

Thus we have defined multivariate convergence in probability to a constant in terms of univariate convergence in probability to a constant.

Multivariate Convergence in Probability (cont.)

Since $\|\mathbf{X}_n - \mathbf{a}\|^2 = \sum_{i=1}^{k} (X_{ni} - a_i)^2$, and

$$(X_{ni} - a_i)^2 \leqslant \sum_{i=1}^{k} (X_{ni} - a_i)^2 = \|\mathbf{X}_n - \mathbf{a}\|^2$$

it follows that

$$\mathbf{X}_n \xrightarrow{P} \mathbf{a}$$

implies

$$X_{ni} \xrightarrow{P} a_i, \quad i = 1, \ldots, k$$

Thus, joint convergence in probability to a constant (of random vectors) implies marginal convergence in probability to a constant (of each component of those random vectors).

Multivariate Convergence in Probability (cont.)

Conversely, if we have

$$X_{ni} \xrightarrow{P} a_i, \quad i = 1, \ldots, k$$

then the continuous mapping theorem implies

$$(X_{ni} - a_i)^2 \xrightarrow{P} 0, \quad i = 1, \ldots, k$$

and Slutsky's theorem implies

$$(X_{n1} - a_1)^2 + (X_{n2} - a_2)^2 \xrightarrow{P} 0$$

So by mathematical induction,

$$||\mathbf{X}_n - \mathbf{a}||^2 \xrightarrow{P} 0$$

In words: joint convergence in probability to a constant (of random vectors) is equivalent to marginal convergence in probability to a constant (of each component of those random vectors).
But multivariate convergence in distribution is different!

Multivariate Convergence in Distribution

If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is a sequence of k-dimensional random vectors, and $\mathbf{X}$ is another k-dimensional random vector, then we say $\mathbf{X}_n$ converges in distribution to $\mathbf{X}$ if

$$\mathbb{E}(g(\mathbf{X}_n) \to E(g(\mathbf{X}))$$

for all bounded continuous functions $g : \mathbb{R}^k \to \mathbb{R}$, and we write

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}$$

Cramèr-Wold Characterization

The Cramèr-Wold theorem asserts that the following is an equivalent characterization of multivariate convergence in distribution:

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}$$

if and only if

$$\mathbf{a}^T\mathbf{X}_n \xrightarrow{d} \mathbf{a}^T\mathbf{X}$$

for every constant conformable vector $\mathbf{a}$.

Multivariate Convergence in Distribution (cont.)

- We have defined multivariate convergence in distribution in terms of univariate convergence in distribution.
- If we use vectors having only one nonzero component in the Cramèr-Wold theorem, we see that joint convergence in distribution (of random vectors) implies marginal convergence in distribution (of each component of those random vectors).
- But the converse is not, in general, true!

Multivariate Convergence in Distribution (cont.)

A simple example where marginal convergence in distribution holds but joint convergence in distribution fails.
Define

$$\mathbf{X}_n = \begin{pmatrix} X_{n1} \\ X_{n2} \end{pmatrix} \quad X_{n1} \sim \mathcal{N}(0,1), X_{n2} = (-1)^n X_{n1}$$

That is, X_{n2} is also standard normal for all n.
Trivially,

$$X_{ni} \xrightarrow{d} \mathcal{N}(0,1), \quad i = 1,2$$

so we have marginal convergence in distribution.
But, if $\mathbf{a} = (1,1)^T$,

$$\mathbf{a}^T\mathbf{X} = \begin{pmatrix} 1 & 1 \end{pmatrix} \begin{pmatrix} X_{n1} \\ X_{n2} \end{pmatrix} = X_{n1}(1 + (-1)^n) = \begin{cases} 2X_{n1}, & n \text{ even} \\ 0 & n \text{ odd} \end{cases}$$

and this sequence does not converge in distribution, so we do not have joint convergence in distribution, that is,

$$\mathbf{X}_n \xrightarrow{d} \mathbf{Y}$$

cannot hold for any random vector $\mathbf{Y}$.

Multivariate Convergence in Distribution (cont.)

- joint convergence in distribution (of random vectors) implies but is not implied by marginal convergence in distribution (of each component of those random vectors).
- There is one special case where marginal convergence in distribution implies joint convergence in distribution: when the components of the random vectors are independent.
  Suppose

$$X_{ni} \xrightarrow{d} Y_i, \quad i = 1, \ldots, k$$

  where X_{ni} is independet of X_{nj} for all $i \neq j$. Then,

$$\mathbf{X}_n \xrightarrow{d} \mathbf{Y}$$

  where $\mathbf{X}_n = (X_{n1}, \ldots, X_{nk})^T$, $\mathbf{Y} = (Y_1, \ldots, Y_k)^T$.

The Multivariate Continuous Mapping Theorem

Theorem

Suppose

$$\mathbf{X}_n \xrightarrow{d} \mathbf{X}$$

and g is a function that is continuous on a set A such that

$$\mathbb{P}(\mathbf{X} \in A) = 1$$

Then,

$$g(\mathbf{X}_n) \xrightarrow{d} g(\mathbf{X})$$

Multivariate Slutsky's Theorem

Suppose

$$\mathbf{X}_n = \begin{pmatrix} \mathbf{X}_{n1} \\ \mathbf{X}_{n2} \end{pmatrix}$$

are partitioned random vectors and

$$\mathbf{X}_{n1} \xrightarrow{d} \mathbf{Y}$$

$$\mathbf{X}_{n2} \xrightarrow{P} \mathbf{a}$$

where $\mathbf{Y}$ is a random vector and $\mathbf{a}$ is a constant vector. Then,

$$\mathbf{X}_n \xrightarrow{d} \begin{pmatrix} \mathbf{Y} \\ \mathbf{a} \end{pmatrix}$$

The constant random vector $\mathbf{a}$ is necessarily independent of the random vector $\mathbf{Y}$, because a constant random vector is independent of any other random vector. Thus there is only one distribution the partitioned random vector

$$\begin{pmatrix} \mathbf{Y} \\ \mathbf{a} \end{pmatrix}$$

can have.

Multivariate Slutsky's Theorem

In conjunction with the continuous mapping theorem, this more general version of Slutsky's theorem implies the earlier version.
For any function g that is continuous at points of the form

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{a} \end{pmatrix}$$

we have

$$g(\mathbf{X}_{n1}, \mathbf{X}_{n2}) \xrightarrow{d} g(\mathbf{Y}, \mathbf{a})$$

The Multivariate CLT

If $\mathbf{X}_1, \mathbf{X}_2, \ldots$ is an iid sequence of k-dimensional random vectors, with mean vector $\boldsymbol{\mu}$ and variance matrix $\boldsymbol{\Sigma}$ and

$$\overline{\mathbf{X}}_n = \frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_i$$

Then,

$$\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$$

- The multivariate CLT follows from the univariate CLT and the Cramèr-Wold theorem.

$$\mathbf{a}^T \left(\sqrt{n}(\overline{\mathbf{X}}_n - \boldsymbol{\mu}) \right) = \sqrt{n}(\mathbf{a}^T \overline{\mathbf{X}}_n - \mathbf{a}^T \boldsymbol{\mu}) \xrightarrow{d} \mathcal{N}(0, \mathbf{a}^T \boldsymbol{\Sigma} \mathbf{a})$$

Multivariate Differentiation

A function $g : \mathbb{R}^d \to \mathbb{R}^k$ is differentiable at a point $\mathbf{x}$ if there exists a matrix $\mathbf{B}$ such that

$$g(\mathbf{x} + \mathbf{h}) = g(\mathbf{x}) + \mathbf{B}\mathbf{h} + o(||\mathbf{h}||)$$

in which case the matrix $\mathbf{B}$ is unique and is called the derivative of the function g at the point $\mathbf{x}$ and is denoted $\nabla g(\mathbf{x})$.

Multivariate Differentiation (cont.)

A sufficient but not necessary condition for the function

$$\mathbf{x} = (x_1, \ldots, x_d) \to g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_k(\mathbf{x}))$$

to be differentiable at a point $\mathbf{x}$ is that all of the partial derivatives $\partial g_i(\mathbf{x})/\partial x_j$ exist and are continuous in which case

$$\nabla g(\mathbf{x}) = \begin{pmatrix} \frac{\partial g_1(\mathbf{x})}{\partial x_1} & \frac{\partial g_1(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_1(\mathbf{x})}{\partial x_d} \\ \frac{\partial g_2(\mathbf{x})}{\partial x_1} & \frac{\partial g_2(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_2(\mathbf{x})}{\partial x_d} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial g_k(\mathbf{x})}{\partial x_1} & \frac{\partial g_k(\mathbf{x})}{\partial x_2} & \cdots & \frac{\partial g_k(\mathbf{x})}{\partial x_d} \end{pmatrix} : k \times d$$

$\nabla g(\mathbf{x})$ is the matrix whose determinant is the Jacobian determinant in the multivariate change-of-variable formula. For this reason it is sometimes called the Jacobian matrix.

The Multivariate Delta Method

The multivariate delta method is just like the univariate delta method. The proofs are analogous.
Suppose

$$n^{\alpha}(\mathbf{X}_n - \theta) \xrightarrow{d} \mathbf{Y}$$

where $\alpha > 0$ and g s a differentiable function at θ. Then,

$$n^{\alpha}(g(\mathbf{X}_n) - g(\theta)) \xrightarrow{d} \nabla g(\mathbf{x})\mathbf{Y}$$

$\nabla g(\mathbf{x})$ is the matrix whose determinant is the Jacobian determinant in the multivariate change-of-variable formula. For this reason it is sometimes called the Jacobian matrix.

The Multivariate Delta Method (cont.)

Since we routinely use the delta method in the case where the rate is $\sqrt{n}$ and the limiting distribution is normal, it is worthwhile working out details of that case.
Suppose

$$\sqrt{n}(\mathbf{X}_n - \boldsymbol{\theta}) \xrightarrow{d} \mathcal{N}(0, \boldsymbol{\Sigma})$$

and suppose g is a function differentiable at $\boldsymbol{\theta}$, then the delta method says

$$\sqrt{n}(g(\mathbf{X}_n) - g(\boldsymbol{\theta})) \xrightarrow{d} \mathcal{N}(0, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$$

where $\mathbf{B} = \nabla g(\mathbf{x})$.

The Delta Method (contd.)

The case where $g : \mathbb{R}^n \to \mathbb{R}$: Let $X_1, \ldots, X_n$ be random variables with expectations $\theta_1, \ldots \theta_n$.
Define $\mathbf{X} = (X_1, \ldots, X_n)$ and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$.
Assume there exists a differentiable function $g(\mathbf{X})$ (an estimator of some parameter) for which we want an approximate estimate of variance.
Let

$$g_i'(\boldsymbol{\theta}) = \frac{\partial}{\partial x_i} g(\mathbf{x}) \Big|_{\mathbf{x}=\boldsymbol{\theta}}$$

The first order Taylor series expansion of g around $\boldsymbol{\theta}$ is

$$g(\mathbf{x}) \approx g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta}) \, (x_i - \theta_i) \tag{1}$$

The Delta Method (contd.)

By taking expectations on both sides of (1) (noticing that everything but the X_i terms on the right-hand side are non-random) we obtain

$$\mathbb{E}\, g(\mathbf{X}) = \mathbb{E}\, g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\, \mathbb{E}(X_i - \theta_i) = g(\boldsymbol{\theta}).$$

We approximate the variance of $g(\mathbf{X})$ as

$$\begin{aligned} Var\, g(\mathbf{X}) &\approx Var\left(g(\boldsymbol{\theta}) + \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\,(X_i - \theta_i)\right) = \mathbb{V}\mathrm{ar}\Big(\sum_{i=1}^{k} g_i'(\boldsymbol{\theta})\,(X_i - \theta_i)\Big) \\ &= \sum_{i=1}^{k} g_i'(\boldsymbol{\theta})^2\, \mathbb{V}\mathrm{ar}(X_i) + 2\sum_{i>j} g_i'(\boldsymbol{\theta})\, g_j'(\boldsymbol{\theta})\, \mathrm{Cov}(X_i, X_j) \end{aligned} \tag{2}$$

where the last equality derives the definition and properties of variance and covariance.
We have approximated the variance to our estimator $g(\mathbf{X})$ using only the variances and covariances of the X_i, which are usually not very difficult to compute or estimate. Independence is not required.

Empirical Distribution

1. Suppose we consider $S = \{1, 2, \ldots, n\}$ as the sample space and X as a random variable having values $X(i) = x_i$.
2. If we consider a uniform distribution on the sample space, i.e. each of the n points, $x_1, \ldots, x_n$, has probability $1/n$, then the distribution of X is called the empirical distribution associated with the vector $\mathbf{x} = (x_1, \ldots, x_n)$.
3. The prob. mass function of X is

$$p(x) = \mathbb{P}(X = x) = \sum_{i \in S} \frac{I(x_i = x)}{n} = \frac{\#(\{i \in S : x_i = x\})}{n}$$

   If all x_i are distinct, then the distribution of X is also uniform.
4. The empirical probability measure associated with the vector $\mathbf{x} = (x_1, \ldots, x_n)$ is denoted P_n and defined by

$$P_n(A) = \frac{1}{n} \sum_{i=1}^{n} I_A(x_i)$$

5. The empirical cdf is defined by

$$F_n(x) = \sum_{i=1}^{n} \frac{I(x_i \leqslant x)}{n} = \frac{\#(\{i \in S : x_i \leqslant x\})}{n}$$

Quantiles

Definition

For $0 < p < 1$, a point x is a p-th quantile of the distribution of a real valued r.v. X if

$$\mathbb{P}(X \leqslant x) \geqslant p \quad \text{and} \quad \mathbb{P}(X \geqslant x) \geqslant 1 - p$$

1. If the c. d. f. of X is invertible: For $0 < p < 1$, the p-th quantile is the unique solution x of the equation

$$F(x) = p, \quad \text{or,} \quad x = F^{-1}(p)$$

2. Fact: A continuous random variable with a strictly positive pdf has an invertible cdf.
3. In general, the p-th quantile need not be unique and it need not be a point satisfying $F(x) = p$.
4. A *usable* definition: A point x is a p-th quantile of a random variable with cdf F if

$$F(x) \geqslant p \quad \text{and} \quad F(y) \leqslant p, \ \forall y < x$$

Quantiles of the Empirical Distribution

1. We denote the sorted values of the components of $\mathbf{x} = (x_1, \ldots, x_n)$ by

$$x_{(1)} \leqslant x_{(2)} \leqslant \ldots \leqslant x_{(n)}$$

2. If np is not an integer, then the p-th quantile of the empirical distribution associated with the vector $\mathbf{x} = (x_1, \ldots, x_n)$ is unique and is equal to $x_{(\lceil np \rceil)}$ ($\lceil z \rceil$ is the smallest integer greater than or equal to z).
3. If np is an integer, then any point x such that

$$x_{(np)} \leqslant x \leqslant x_{(np+1)}$$

is a p-th quantile.

Quantiles of the Empirical Distribution

1. **Example:** Suppose we have the 10 sorted points 0.08 0.12 0.29 0.35 0.49 0.77 0.81 1.02 1.05 3.15
   1. 0.25 quantile: $np = 10 \times 0.25 = 2.5$ with $\lceil 2.5 \rceil = 3$, thus 0.25 quantile is the third observation: 0.29
   2. 0.40 quantile: $np = 10 \times 0.4 = 4$. Then, any point between $x_{(4)} = 0.35$ and $x_{(5)} = 0.49$ is a 0.40 quantile.
2. In R

```
x=c(0.08,  0.12 , 0.29,  0.35 , 0.49 , 0.77,
0.81 , 1.02 , 1.05 , 3.15)
plot(ecdf(x)) # plots the empirical cdf

quantile(x, type = 1) #calculates several quantiles
quantile(x, probs = 1 / 3, type = 1)
```

Empirical Median

Definition

The median of the values $x_1, \ldots, x_n$ is the middle value in sorted order when n is odd, $x_{(\lceil n/2 \rceil)}$ and

$$\frac{x_{(n/2)} + x_{(n/2+1)}}{2}$$

when n is even.

1. In our example, the median is

$$\frac{x_{(5)} + x_{(6)}}{2} = \frac{0.49 + 0.77}{2} = 0.63$$

2. in R: `median(x)`

Exercise: If X is a real-valued random variable with finite expectation, then a median of X is any value of a that minimizes the function $g(a) = \mathbb{E}(|X - a|)$. Also, a median of the empirical distribution is a value of a that minimizes the function $g(a) = \sum_{i=1}^{n} |x_i - a|/n$.

Random Sampling

1. A population is any set of subjects.
2. A sample is any subset of the population.
3. A random sample is one drawn so that every member of the population is equally likely to be in the sample. There are two types.
    1. Sampling without Replacement: $\binom{N}{n}$ possible samples of size n from a population of size N, all equally likely
    2. Sampling with Replacement: N^n possible samples of size n from a population of size N, all equally likely
4. When we take a random sample of size n from the population we obtain a sequence $X_1, \ldots, X_n$ of values of the variable X whose values form the population. Each X_i is one of the population values x, but it randomly takes this value.

Sampling Distributions

If $X_1, \ldots, X_n$ are a random sample from a population of size N, then the marginal distribution of each X_i is the empirical distribution associated with the population values $x_1, \ldots, x_N$. If the sampling is with replacement, then the X_i are independent and identically distributed. If the sampling is without replacement, then the X_i are exchangeable but not independent.

1. The X_is are exchangeable by definition: every permutation of the sample is equally likely. Hence they are identically distributed, and the marginal distribution of X_i is the marginal distribution of X_1.
2. Since every subject is equally likely to be the first one drawn, X_1 has the empirical distribution.
3. Under sampling with replacement, every sample has probability $1/N^n$, which is the product of the marginals, i.e., X_is are independent.
4. Under sampling without replacement, every sample has probability $1/\binom{N}{n}$, which is not the product of the marginals, i.e., X_is are dependent.

Random Samples

Definition

The random variables $X_1, \ldots, X_n$ are called a *random sample of size n from the population* $f(x)$ ($p(x)$) if $X_1, \ldots, X_n$ are mutually independent and identically distributed (i.i.d.) random variables with pdf $f(x)$ (pmf $p(x)$).

1. When the population is finite ($N < \infty$), we focus on *sampling with replacement*
2. The random sampling model describes a type of experimental situation in which the variable of interest (population) has a probability distribution described by $f(x)$ ($p(x)$). In most experiments there are $n > 1$ repeated observations made on the variable of interest. The observations are taken in such way that the value of one observation has *no* effect on the relationship with any of other observations (mutually independent).

Random Samples

Important formulas:

1. the variance of a sum is the sum of the variances

$$\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i} \mathbb{V}\mathrm{ar}(X_i) = n\sigma^2$$

2. The joint pdf of $(X_1, \ldots, X_n)$ is the product of the correspoding marginal pdfs, i.e.

$$f(x_1, \ldots, x_n) = f_{X_1}(x_1) \cdot \cdots \cdot f_{X_n}(x_n) = \prod_{i=1}^{n} f_{X_i}(x_i),$$

   since $X_1, \ldots, X_n$ are identically distributed, their marginal pdfs are the same function $f(x)$, i.e.
3. For sampling without replacement neither of the above holds.

Random Samples

1. In particular, if the population pdf is a member of a parametric family with pdf given by $f(x, \theta)$, then the joint pdf is

$$f(x_1, \ldots, x_n, \theta) = \prod_{i=1}^{n} f(x_i, \theta), \tag{3}$$

   where the same value of the parameter θ is used in each of the terms in the product.
2. If we assume that the population we are observing is a member of a specific parametric family, but the *true parameter value is unknown*, then a random sample from this population has a joint pdf of the form (3) and the sample is used to estimate θ.

Random Samples: Sample pdf exponential

Consider the times $X_1, \ldots, X_n$ (in years) until failure of n identical circuit boards that are put on test and used until they fail. Then we can assume $X_1, \ldots, X_n$ to be a random sample from an $Exp(\lambda)$ population.
The goal is to compute the probability that all the circuit boards last more than two years. One way to obtain this probability is to use the joint pdf

$$f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x, \lambda) = \prod_{i=1}^{n} \left(\lambda \cdot e^{-\theta x_i} \right) = \lambda^n \cdot e^{-\lambda (x_1 + \cdots + x_n)}$$

for all $x_i \geqslant 0$, $i = 1, ..., n$. Then, the unknown probability equals

$$P(X_1 > 2, \ldots, X_n > 2) = \int_2^{+\infty} \cdots \int_2^{+\infty} \lambda^n \cdot e^{-\lambda (x_1 + \cdots + x_n)} dx_n \ldots dx_1 = \cdots = e^{-2\lambda n},$$

Another way is to calculate it directly, by using the i.i.d. assumption:

$$\begin{aligned} P(X_1 > 2, \ldots, X_n > 2) &= P(X_1 > 2) \cdot \ldots P(X_n > 2) \\ &= \Big(P(X_1 > 2) \Big)^n \\ &= \Big(\int_2^{+\infty} \lambda e^{-\lambda x} \, dx \Big)^n = \left(e^{-2\lambda} \right)^n = e^{-2\lambda n}. \end{aligned}$$

Random Samples: Finite Population Correction

Exercise:
If $X_1, \ldots, X_n$ are a random sample without replacement from a finite population of size N, then all X_is have the same variance σ^2 and

$$\mathbb{V}\mathrm{ar}\left(\sum_{i=1}^{n} X_i\right) = n\sigma^2 \, \frac{N-n}{N-1}$$

The factor $\frac{N-n}{N-1}$ is called the finite population correction.

Statistics and sampling distributions

Definition

Let $X_1, \ldots, X_n$ be a random sample of size n from a population with parameter θ. Also, let $T(X_1, \ldots, X_n)$ be a real-valued or vector-valued function whose domain includes the sample space of $(X_1, \ldots, X_n)$. Then the random variable, or, random vector

$$Y = T(X_1, \ldots, X_n)$$

is called a *statistic*. The probability distribution of a statistic Y is called the *sampling distribution* of Y.

The only restriction in the definition of statistic is that it cannot be a function of a parameter θ.

Sample Moments

If $X_1, \ldots, X_n$ are a random sample, the sample moments are the moments of the empirical distribution associated with the vector $\mathbf{X} = (X_1, \ldots, X_n)$.

1. The first moment is the *sample mean*, i.e. the arithmetic average of the values in a random sample

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{4}$$

2. The k-th sample moment is

$$\frac{1}{n} \sum_{i=1}^{n} X_i^k$$

3. The k-th central moment is

$$\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X}_n)^k$$

Sample Moments

1. The second central moment or *modified sample variance* is

$$\tilde{S}_n^2 = \frac{1}{n} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2$$

2. The *sample variance* is the statistic defined by

$$S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X} \right)^2. \tag{5}$$

3. The *sample standard deviation* is the statistic defined by

$$S_n = \sqrt{S_n^2}.$$

Sample mean distribution

1. All sample moments are random variables, so they have a probability distribution
2. In some cases we know the distribution of $\bar{X}_n$.
3. Let $Y = \sum_{i=1}^{n} X_i = n\bar{X}_n$.
4. We know the distribution of Y when X_i iid
    1. If $X_i \sim Bernoulli(p)$, then $n\bar{X}_n \sim Bin(n, p)$
    2. If $X_i \sim Geo(p)$, then $n\bar{X}_n \sim NegBin(n, p)$
    3. If $X_i \sim Poi(\lambda)$, then $n\bar{X}_n \sim Poi(n\lambda)$ Poisson
    4. If $X_i \sim Exp(\lambda)$, then $n\bar{X}_n \sim Gamma(n, \lambda)$
    5. If $X_i \sim \mathcal{N}(\mu, \sigma^2)$, then $n\bar{X}_n \sim cN(n\mu, n\sigma^2)$
5. If f_Y is the distribution of Y, and X_i are continuous r.v.'s,

$$f_{\bar{X}_n} = nf_Y(nz)$$

Example: Exponential

1. Let $X_1, \ldots, X_n$ be iid $Exp(\lambda)$.
2. Also, $Exp(\lambda)$ is Gamma$(1, \lambda)$.
3. Then, $Y = n\bar{X}_n \sim$ Gamma(n, λ)
4. Thus,
$$\bar{X}_n \sim Gamma(n, n\lambda)$$
5. If $X \sim Gamma(n, \lambda)$, then
$$2\lambda X \sim \chi^2(2n)$$
This allows one to use the chi-squared distn tables.
6. **Show the above!**

Asymptotic Distributions of Sample Moments

Typically we cannot calculate the exact sampling distribution of a sample moment, but we can always get large sample properties of the distribution from the law of large numbers, the central limit theorem, and Slutsky's theorem.

1. Under i. i. d. sampling every sample moment converges in probability to the corresponding population moment provided the population moment exists.
2. For example,
$$\frac{1}{n}\sum_{i=1}^{n} X_i^k \xrightarrow{P} \mathbb{E}(X_1^k)$$
3. Also, If each X_i has mean μ and variance σ^2, then $\bar{X}_n$ is approximately $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$

Asymptotic Distributions of Sample Moments

1. $$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{k}\binom{k}{j}(-1)^j(\bar{X}_n-\mu)^j(X_i-\mu)^{k-j}$$
$$= \sum_{j=0}^{k}\binom{k}{j}(-1)^j(\bar{X}_n-\mu)^j\frac{1}{n}\sum_{i=1}^{n}(X_i-\mu)^{k-j}$$

   by the LLN, the continuous mapping theorem and Slutsky's,

$$\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X}_n)^k \xrightarrow{P} \mathbb{E}(X_1-\mu)^k$$

2. Suppose $X_1,\ldots,X_n$ are i. i. d. and have fourth moments. Then,

$$\sqrt{n}\left(\frac{n-1}{n}S_n^2-\sigma^2\right) \xrightarrow{d} \mathcal{N}(0,\mu_4-\mu_2^2)$$

   where μ_2, μ_4 are the 2nd and 4th central moments.

Important Distributions

1. Chi-Square Distribution: for any real number $\nu > 0$, the chi-square distribution with ν degrees of freedom, abbreviated $\chi^2(\nu)$, is the Gamma$\left(\frac{\nu}{2}, \frac{1}{2}\right)$ distribution.
2. its pdf is given by

$$f(x) = \frac{1}{\Gamma\left(\frac{\nu}{2}\right) 2^{\frac{\nu}{2}}} \cdot x^{\frac{\nu}{2}-1} \cdot e^{-\frac{x}{2}}, \quad x > 0.$$

3. A useful result for computing moments of a chi squared distribution: For the proof we refer to [1]. Let $X \sim \chi^2(\nu)$. For any function $h(x)$ it holds

$$\mathbb{E}(h(X)) = \nu \, E\left(\frac{h(Y)}{Y}\right), \tag{6}$$

provided the expectations exist, where $Y \sim \chi^2(\nu + 2)$. By using (6) we compute the expectation and variance of $X \sim \chi^2(\nu)$,

$$\mathbb{E}(X) = \nu \qquad \text{and} \qquad \mathbb{V}\text{ar}(X) = 2\nu.$$

Important Distributions

1. Student's t Distribution: Suppose Z and Y are independent random variables with

$$Z \sim \mathcal{N}(0,1)$$
$$Y \sim \chi^2(\nu)$$

Then,

$$t = \frac{Z}{\sqrt{Y/\nu}}$$

is said to have Student's t distribution with ν degrees of freedom, abbreviated $t(\nu)$.

2. **Exercise:** Derive the pdf of a $t(\nu)$ random variable

$$f_\nu(x) = \frac{1}{\sqrt{\nu\pi}} \cdot \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \cdot \frac{1}{\left(1+\frac{x^2}{\nu}\right)^{(\nu+1)/2}}, \quad x \in \mathbb{R}$$

Moments of $t(\nu)$

1. For the $t(\nu)$ distribution and $k > 0$,

$$\mathbb{E}(|X|^k) \quad \text{exists if and only if} \quad k < \nu$$

Proof: The pdf is bounded, so we only need to examine the behavior of the pdf at $\pm\infty$. Since the t distribution is symmetric about zero, we only need to check the behavior at ∞. When does

$$\int_0^\infty x^k f_\nu(x) dx$$

exist?

$$\lim_{x\to\infty} \frac{x^k f_\nu(x)}{x^\alpha} \to c$$

when $\alpha = k - (\nu + 1)$ so that the integral exists iif

$$k - (\nu + 1) < -1$$

which is equivalent to $k < \nu$.

Moments of $t(\nu)$

1. The $t(\nu)$ distribution is symmetric about zero, hence the mean is zero if it exists. Also, central moments are equal to ordinary moments and every odd ordinary moment is zero if it exists:
   1. If $X \sim t(\nu)$ and $\nu > 1$, then
   $$\mathbb{E}(X) = 0$$
   by symmetry. If $\nu \leqslant 1$, the mean does not exist!
   2. If $X \sim t(\nu)$ and $\nu > 2$, then
   $$\mathbb{V}\text{ar}(X) = \frac{\nu}{\nu - 2}$$
   otherwise the variance does not exist (**exercise**)

The Cauchy distribution

1. Plugging in $\nu = 1$ into the formula for the pdf of the $t(\nu)$ distribution gives the pdf of the standard Cauchy distribution:

$$t(1) = Cauchy(0, 1)$$

2. That is, if Z_1 and Z_2 are iid $\mathcal{N}(0, 1)$, then

$$T = \frac{Z_1}{Z_2} \sim Cauchy(0, 1)$$

t and Standard Normal Distribution

1. If $Y \sim \chi^2(\nu) = Gamma(\nu/2, 1/2)$, then $U = Y/\nu \sim Gamma(\nu/2, \nu/2)$, so that

$$\mathbb{E}(U) = \frac{\nu}{2}\frac{2}{\nu} = 1$$

$$\mathbb{V}\mathrm{ar}(U) = \frac{\nu}{2}\frac{2^2}{\nu^2} = \frac{2}{\nu}$$

2. That is,

$$U \xrightarrow{P} 1, \quad \text{as } \nu \to \infty$$

   by Chebyshev's inequality.

3. Hence if Z is a standard normal random variable independent of Y,

$$\frac{Z}{\sqrt{Y/\nu}} \xrightarrow{d} Z, \quad \text{as } \nu \to \infty$$

   by Slutsky's theorem.

4. We showed that

$$t(\nu) \xrightarrow{d} \mathcal{N}(0,1), \quad \text{as } \nu \to \infty$$

F Distribution

1. If X and Y are independent random variables and

$$X \sim \chi^2(\nu_1)$$

$$Y \sim \chi^2(\nu_2)$$

then

$$F = \frac{X/\nu_1}{Y/\nu_2}$$

has the F distribution with ν_1 numerator degrees of freedom and ν_2 denominator degrees of freedom

2. This random variable was introduced by Snedecor and named after R.A. Fisher. It is abbreviated as $F(\nu_1, \nu_2)$.

3. Using similar arguments as to obtain $t(\nu) \xrightarrow{d} \mathcal{N}(0,1)$, as $\nu \to \infty$, we can obtain

$$F(\nu_1, \nu_2) \xrightarrow{P} 1, \quad \text{as} \;\; \nu_1, \nu_2 \to \infty$$

F and Beta Distribution

1. In an exercise you had shown that if $X \sim Gamma(\alpha_1, \lambda)$ and $Y \sim Gamma(\alpha_2, \lambda)$ are independent random variables, then
$$V = \frac{X}{X+Y} \sim Beta(\alpha_1, \alpha_2)$$
2. Alternatively, if $X \sim \chi^2(\nu_1)$ and $Y \sim \chi^2(\nu_2)$ are independent random variables, then
$$V = \frac{X}{X+Y} \sim Beta\left(\nu_1/2, \nu_2/2\right)$$

F and Beta Distribution

1. Since
$$\frac{X}{Y} = \frac{V}{1-V}$$
$$W = \frac{\nu_2}{\nu_1} \cdot \frac{V}{1-V}$$
and
$$V = \frac{\nu_1 W/\nu_2}{1+\nu_1 W/\nu_2}$$
2. which gives the relationship between $W \sim F(\nu_1, \nu_2)$ and $V \sim Beta\,(\nu_1/2, \nu_2/2)$

Sampling Distributions for Normal Populations

Sampling from a population with normal distribution leads to many useful properties of sample statistics and also many well-known sampling distributions (such as χ^2, t and F distributions).
Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ distribution. Then,

(a) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$

(b) $\bar{X}$ and S^2 are *independent* random variables.

(c) $(n-1)\frac{S^2}{\sigma^2} \sim \chi^2_{n-1}$, chi squared distribution with $(n-1)$ degrees of freedom.

Proof

(b) Without a loss of generality, assume $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(0,1)$. First, represent S^2 in the form

$$S^2 = \frac{1}{n-1}\left(\Big(\sum_{i=2}^{n} (X_i - \bar{X}) \Big)^2 + \sum_{i=2}^{n} \big(X_i - \bar{X}\big)^2 \right),$$

i.e. as a function only of $(X_2 - \bar{X}, \ldots, X_n - \bar{X})$.
Then, in order to show that these random variables are independent of $\bar{X}$, one has to show that they are uncorrelated with $\bar{X}$, i.e.
$\mathbb{C}\mathrm{ov}(\bar{X}, X_j - \bar{X}) = 0$, for all $j = 2, \ldots, n$.

Proof

$$
\begin{aligned}
\mathbb{C}\mathrm{ov}(\bar{X}_n, X_i - \bar{X}_n) &= \mathbb{C}\mathrm{ov}(\bar{X}_n, X_i) - \mathbb{V}\mathrm{ar}(\bar{X}_n) \\
&= \mathbb{C}\mathrm{ov}(\bar{X}_n, X_i) - \frac{\sigma^2}{n} \\
&= \mathbb{C}\mathrm{ov}\left(\frac{1}{n}\sum_{j=1}^{n} X_j, X_i\right) - \frac{\sigma^2}{n} \\
&= \frac{1}{n}\sum_{j=1}^{n} \mathbb{C}\mathrm{ov}(X_j, X_i) - \frac{\sigma^2}{n} \\
&= 0
\end{aligned}
$$

Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ distribution. Then,

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

$$\frac{(n-1)S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

Hence,

$$t = \frac{\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{\frac{(n-1)S_n^2}{\sigma^2}}{n-1}}} = \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}} \sim t(n-1)$$

Theorem

Let $x_1, \ldots, x_n$ be any numbers. Consider also the numbers $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2$. Then the following properties hold

(a) $\min_a \sum_{i=1}^{n} (x_i - a)^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2$

(b) $(n-1)\, s^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - n\bar{x}^2.$

Proof

(a) We first add and substract $\bar{x}$.

$$\begin{aligned}
\sum_{i=1}^{n}(x_i-a)^2 &= \sum_{i=1}^{n}(x_i-\bar{x}+\bar{x}-a)^2 \\
&= \sum_{i=1}^{n}\Big((x_i-\bar{x})^2+2(x_i-\bar{x})(\bar{x}-a)+(\bar{x}-a)^2\Big) \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2+2(\bar{x}-a)\sum_{i=1}^{n}(x_i-\bar{x})+n(\bar{x}-a)^2 \\
&= \sum_{i=1}^{n}(x_i-\bar{x})^2+n(\bar{x}-a)^2
\end{aligned} \tag{7}$$

The middle term in the last step equals zero, since $\sum_{i=1}^{n}(x_i-\bar{x}) = \sum_{i=1}^{n}x_i - n\bar{x} = 0$. It is clear that the right hand side of (7) is minimized at $a=\bar{x}$.

(b) To prove (b) we take $a=0$ in (7). Then,

$(n-1)s^2 = \sum_{i=1}^{n}x_i^2 = \sum_{i=1}^{n}(x_i-\bar{x})^2 + n\bar{x}^2$. □

Properties of Sample Moments

Lemma

Let $X_1, \ldots, X_n$ be a random sample from a population and let $g(x)$ be a function such that $\mathbb{E}(g(X_1))$ and $\mathbb{V}\mathrm{ar}\,(g(X_1))$ exist. Then,

$$\mathbb{E}\Big(\sum_{i=1}^{n} g(X_i)\Big) = n\,\mathbb{E}\,g(X_1) \quad \text{and} \quad \mathbb{V}\mathrm{ar}\,\Big(\sum_{i=1}^{n} g(X_1)\Big) = n\,\mathbb{V}\mathrm{ar}\,g(X_1)$$

The proof is left for HW. More details can be found in [1].

Theorem

Let $X_1, \ldots, X_n$ be arandom sample from a population with expectation μ and variance $\sigma^2 < \infty$. Then,

(a) $\mathbb{E}(\bar{X}_n) = \mu$
(b) $\mathbb{V}\mathrm{ar}(\bar{X}_n) = \frac{\sigma^2}{n}$
(c) $\mathbb{E}(S_n^2) = \sigma^2$.

Proof.

Although, in our previous lectures, we obtained the forms (a) and (b) of the expectation and the variance of the sample mean, here we repeat their proofs.

(a) $\mathbb{E}(\bar{X}_n) = \mathbb{E}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n}\cdot n\mathbb{E}(X_1) = \mu.$

(b) $\mathbb{V}\mathrm{ar}(\bar{X}_n) = \mathbb{V}\mathrm{ar}\Big(\frac{1}{n}\sum_{i=1}^{n} X_i\Big) = \frac{1}{n^2}\cdot\mathbb{V}\mathrm{ar}(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\cdot n\mathbb{V}\mathrm{ar}\, X_1 = \frac{\sigma^2}{n}.$

(c) $\mathbb{E}(S_n^2) = \mathbb{E}\Big(\frac{1}{n-1}\cdot(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)\Big) = \frac{1}{n-1}\cdot\Big(n\mathbb{E}(X_1^2) - n\bar{X}_n^2\Big) =$
$\frac{1}{n-1}\big(n(\sigma^2+\mu^2) - n(\frac{\sigma^2}{n}+\mu^2)\big) = \sigma^2.$

In the final step we represented the second moment in terms of variance, i.e. $\mathbb{E}(X^2) = \mathbb{V}\mathrm{ar}X + (\mathbb{E}(X))^2$. □

Unbiasedness

The relationships (a) and (c), that are relationships between a statistic and a population parameter, are examples of *unbiased statistics*. Namely,

$$\begin{aligned} \bar{X} &\quad \text{is an unbiased statistic of } \mu, \quad \text{i.e.} \quad \mathbb{E}\,(\bar{X}_n) = \mu \quad \text{and} \\ S_n^2 &\quad \text{is an unbiased statistic of } \sigma^2, \quad \text{i.e.} \quad \mathbb{E}(S_n^2) = \sigma^2. \end{aligned}$$

The statistic $\tilde{S}_n^2$ is *not* an unbiased statistic of σ^2, because

$$\mathbb{E}\,\tilde{S}_n^2 = \frac{n-1}{n}\,\sigma^2 \neq \sigma^2$$

- We focus now on the sampling distribution of $\bar{X}_n$
- Let $X_1, \ldots, X_n$ be a random sample from a population with mgf $M_X(t)$. Then, the mgf of the sample mean is

$$\begin{aligned} M_{\bar{X}}(t) &= M_{\frac{1}{n}(X_1+\cdots+X_n)}(t) = M_{X_1+\cdots+X_n}\left(\frac{t}{n}\right) \\ &= M_{X_1}(t) \cdot \ldots \cdot M_{X_n}(t) = \left(M_X\left(\frac{t}{n}\right)\right)^n. \end{aligned}$$

- We used the properties $M_{aX+b}(t) = e^{bt} M_X(at)$ and $M_{X_1+X_2}(t) = M_{X_1}(t) \cdot M_{X_2}(t)$ for X_1 and X_2 independent.

Example

1. Let $X_1, \ldots, X_n$ be a random sample from $\mathcal{N}(\mu, \sigma^2)$ population.
2. We showed that $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.
3. Let $X_1, \ldots, X_n$ be a random sample from $Gamma(\alpha, \beta)$ distribution. Then,

$$\bar{X} \sim Gamma(n\alpha, \frac{\beta}{n}). \tag{8}$$

To show this, we recall that the mgf of $X \sim Gamma\big(\alpha, \beta\big)$ is of the form $\left(\frac{1}{1-\beta t}\right)^{\alpha}$ for $t < \frac{1}{\beta}$. Thus,

$$M_{\bar{X}}(t) = \left[\left(\frac{1}{1-\beta\,\frac{t}{n}}\right)\right]^n = \left(\frac{1}{1-\left(\frac{\beta}{n}\right)t}\right)^{n\alpha},$$

and we conclude that the sample mean follows the gamma distribution (8).

In contrast to the case where the data are exactly normally distributed, in general, $\bar{X}_n$ and S_n^2 are not independent and are not even asymptotically uncorrelated unless the population third central moment is zero (as it would be for any symmetric population distribution but would not be for any skewed population distribution).

Moreover, in general, the asymptotic distribution of S_n^2 is different from what one would get if a normal population distribution were assumed.

Order statistics

Definition

The *order statistics* of a random sample $X_1, \ldots, X_n$ are the sample values placed in ascending order. They are denoted by $X_{(1)}, \ldots, X_{(n)}$.

The order statistics are random variables that satisfy

$$X_{(1)} \leqslant \cdots \leqslant X_{(n)},$$

where

$$X_{(1)} = \min_{1 \leqslant i \leqslant n} X_{(i)} \qquad \text{and} \qquad X_{(n)} = \max_{1 \leqslant i \leqslant n} X_{(i)}.$$

The distance between the smallest and largest observations $R = X_{(n)} - X_{(1)}$ is called the *sample range*.

Order statistics

1. For any number p between 0 and 1 the $(100p)$*th sample percentile* is the observation such that approximately np of the observations are less than this observation and $n(1-p)$ of the observations are greater.
2. Particularly, the 50th sample percentile $p = 0.5$ is the *sample median*, denoted by M.
3. The 25th and 75th sample percentiles are called (*lower and upper*) *sample quartiles* and are respectively denoted by $Q_{0.25}$ and $Q_{0.75}$.
4. The difference $IQR = Q_{0.75} - Q_{0.25}$ is called the (sample) *interquartile range*.

Sampling Distribution of Order Statistics

The cdf of $X_{(k)}$ is

$$\begin{aligned} F_{X_{(k)}}(x) &= \mathbb{P}(X_{(k)} \leqslant x) \\ &= \mathbb{P}(\text{at least } k \text{ of the } X_i \text{ are } \leqslant x) \\ &= \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j} \end{aligned}$$

where $F(x) = \mathbb{P}(X_i \leqslant x)$.

Sampling Distribution of Order Statistics

If the data are continuous random variables with pdf $f = F'$, then the pdf of $X_{(k)}$ is

$$
\begin{aligned}
f_{X_{(k)}}(x) &= F'_{X_{(k)}}(x) \\
&= \frac{d}{dx} \sum_{j=k}^{n} \binom{n}{j} F(x)^j (1 - F(x))^{n-j} \\
&= \sum_{j=k}^{n} \binom{n}{j} j F(x)^j f(x) (1 - F(x))^{n-j} \\
&\quad - \sum_{j=k}^{n-1} \binom{n}{j} F(x)^j (n-j)(1 - F(x))^{n-j-1} f(x)
\end{aligned}
$$

Sampling Distribution of Order Statistics

Replace j by $j-1$ in the second term,

$$
\begin{aligned}
f_{X_{(k)}}(x) &= \sum_{j=k}^{n} \binom{n}{j} jF(x)^j f(x)(1-F(x))^{n-j} \\
&\quad - \sum_{j=k+1}^{n} \binom{n}{j-1} F(x)^{j-1}(n-j+1)(1-F(x))^{n-j} f(x) \\
&= \frac{n!}{(k-1)!(n-k)!} F(x)^{k-1}(1-F(x))^{n-k} f(x)
\end{aligned}
$$

Sampling Distribution of Order Statistics

If $X_1, \ldots, X_n$ are iid $Unif(0,1)$, then the pdf of the kth order statistic is

$$f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!} x^{k-1}(1-x)^{n-k}, \quad 0 < x < 1$$

which is the pdf of a $Beta(k, n-k+1)$ distribution.

Order Statistics

1. **[HW]** Let $X_1, \ldots, X_n$ be a random sample from a continuous population with the pdf $f_X(x)$ and cdf $F_X(x)$. Find the cdfs of the sample minimum $X_{(1)}$ and the sample maximum $X_{(n)}$.
2. Asymptotic distributions of order statistics: we cannot get a normal approximation directly from the CLT because the sum of iid beta distributions does not have a known distribution
3. **[HW]** Can show

$$Beta(\alpha_1, \alpha_2) \approx \mathcal{N}\left(\frac{\alpha_1}{\alpha_1 + \alpha_2}, \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^3}\right)$$

when α_1, α_2 are both large. *Hint:* write

$$W = \frac{X}{X + Y}$$

$X \sim Gamma(\alpha_1, \lambda)$, $Y \sim Gamma(\alpha_2, \lambda)$, independent ad show
$X \approx \mathcal{N}(\alpha_1, \alpha_1)$, $Y \approx \mathcal{N}(\alpha_2, \alpha_2)$

Asymptotic distribution of Order Statistics

Theorem

Suppose $X_1, X_2, \ldots$ are iid r.v.s from a continuous distribution with nonzero pdf f supported on an interval. Let x_p denote the p-th quantile of f. Suppose

$$\sqrt{n}\left(\frac{k_n}{n} - p\right) \to 0, \quad n \to \infty$$

and let Q_n denote the k_n-th order statistic of $X_1, \ldots, X_n$. Then,

$$\sqrt{n}\left(Q_n - x_p\right) \xrightarrow{d} \mathcal{N}\left(0, \frac{p(1-p)}{f(x_p)^2}\right)$$