Samples and Statistics

Paolo Zacchia

Statistics and Econometrics

Lecture 4

Samples

Statistical analysis is based on data **samples** collected from a population of interest.

Definition 1

Sample. A *sample* is a collection $\{\mathbf{x}_i\}_{i=1}^N$ of *realizations* of some N random vectors $\{\boldsymbol{x}_i\}_{i=1}^N$ associated with some population of interest. Each unit of this population is typically called a *unit of observation* and its associated realization $\mathbf{x}_i$ is identified by a unique subscript i.

Note: if the data are collected from N random variables, the sample is written as $\{x_i\}_{i=1}^N$; if from N random matrices, as $\{\mathbf{X}_i\}_{i=1}^N$.

Definition 2

Sample size. The dimension N of a sample is called *size*.

Random samples

Definition 3

Random sample. A sample is *random* if its realizations are drawn from **independent and identically distributed** (**i.i.d.**) random vectors $\{\boldsymbol{x}_i\}_{i=1}^{N}$ (or variables, or matrices).

Random samples are thought to be the product of **sampling with replacement** from a population distributed according to a random vector $\boldsymbol{x}$. Importantly, not all samples are random.

Definition 4

Non-random sample. A sample is *non-random* if the realizations that compose it are not drawn from i.i.d. random variables, vectors, or matrices. Instead, these may be:

- **independent and not identically distributed** (**i.n.i.d.**);
- **not independent and identically distributed** (**n.i.i.d.**);
- **not independent, not identically distributed** (**n.i.n.i.d.**).

Samples and Statistics

While non-random samples are common (they are ubiquitous, say, in econometrics) random samples are an important benchmark. In fact, the i.i.d. property lets express the *joint distribution of the sample* as:

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\theta}\right) = f_{\boldsymbol{x}}\left(\mathbf{x}_1;\boldsymbol{\theta}\right)\times\cdots\times f_{\boldsymbol{x}}\left(\mathbf{x}_N;\boldsymbol{\theta}\right) = \prod_{i=1}^{N} f_{\boldsymbol{x}}\left(\mathbf{x}_i;\boldsymbol{\theta}\right)$$

this is the distribution of an NK-long random vector $(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N)$. It is easier to study the distribution of **statistics** with random samples.

Definition 5

Statistic. A function of the N random variables, vectors or matrices that are specific to each i-th unit of observation and that generate a sample is called a *statistic*. Any statistic is itself a random variable, vector or matrix.

Definition 6

Sampling distribution. The probability distribution of a statistic is called its *sampling* distribution.

Sample mean

Two most common and important sample statistics are defined next.

Definition 7

Sample mean. In samples derived from random vectors, the *sample mean* is a vector-valued statistic usually denoted as $\bar{\boldsymbol{x}}$ and defined as follows.

$$\bar{\boldsymbol{x}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{x}_i$$

This definition can be reduced to samples that drawn from univariate random variables, in which case the usual notation is $\bar{X}$:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

or extended to samples drawn from random matrices, where one can write $\bar{\boldsymbol{X}}$ and the definition is again analogous.

Sample variance-covariance

Definition 8

Sample variance-covariance. In samples collected from random vectors, the *sample variance-covariance* is a matrix-valued statistic usually denoted by $\boldsymbol{S}$ and defined as follows.

$$\boldsymbol{S} = \frac{1}{N-1} \sum_{i=1}^{N} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right) \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}} \right)^{\mathrm{T}}$$

In samples from univariate random variables, this statistic is simply called *sample variance*, its associated notation is S^2, and it is a scalar.

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

In this specific case, the square root of the sample variance is written $S = \sqrt{S^2}$ and is called the *standard deviation*. In order to extend this definition to sampling from random matrices it is necessary to develop three-dimensional arrays.

Properties of key sample statistics, I (1/3)

In what follows, some key results about the sample mean and sample variance-covariance are presented. They are crucial for the derivation moments and sometimes, distribution of these statistics.

These properties are collected in three "cumulative" theorems.

Theorem 1

Properties of simple sample statistics (1). *Consider a sample* $\{\boldsymbol{x}_i\}_{i=1}^N$*, its sample mean* $\bar{\boldsymbol{x}}$*, and its sample variance-covariance* $\boldsymbol{S}$*. The following two properties are true:*

a. $\bar{\boldsymbol{x}} = \arg\min_{\mathbf{a}\in\mathbb{R}^K} \sum_{i=1}^N (\boldsymbol{x}_i - \mathbf{a})(\boldsymbol{x}_i - \mathbf{a})^{\mathrm{T}}$;

b. $(N-1)\,\boldsymbol{S} = \sum_{i=1}^N \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} - N \cdot \bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}$.

Proof.

(Continues...)

Properties of key sample statistics, I (2/3)

Theorem 1

Proof.

(Continued.) To show point **a.** note that:

$$\begin{aligned}
\sum_{i=1}^{N} (\boldsymbol{x}_i - \mathbf{a}) (\boldsymbol{x}_i - \mathbf{a})^{\mathrm{T}} &= \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \mathbf{a}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}} + \bar{\boldsymbol{x}} - \mathbf{a})^{\mathrm{T}} \\
&= \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{i=1}^{N} (\bar{\boldsymbol{x}} - \mathbf{a}) (\bar{\boldsymbol{x}} - \mathbf{a})^{\mathrm{T}} \\
&+ \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\bar{\boldsymbol{x}} - \mathbf{a})^{\mathrm{T}} + \sum_{i=1}^{N} (\bar{\boldsymbol{x}} - \mathbf{a}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} \\
&= \sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} + \sum_{i=1}^{N} (\bar{\boldsymbol{x}} - \mathbf{a}) (\bar{\boldsymbol{x}} - \mathbf{a})^{\mathrm{T}}
\end{aligned}$$

where two terms in the second line are both equal to zero by definition of sample mean; in the last line, the first term does not depend on $\mathbf{a}$ while the second is minimized at $\mathbf{a} = \bar{\boldsymbol{x}}$. **(Continues...)**

Properties of key sample statistics, I (3/3)

Theorem 1

Proof.
(Continued.) To show **b.** simply note that:

$$\sum_{i=1}^{N} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) (\boldsymbol{x}_i - \bar{\boldsymbol{x}})^{\mathrm{T}} = \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} - \sum_{i=1}^{N} \boldsymbol{x}_i \bar{\boldsymbol{x}}^{\mathrm{T}} - \sum_{i=1}^{N} \bar{\boldsymbol{x}} \boldsymbol{x}_i^{\mathrm{T}} + N \cdot \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^{\mathrm{T}}$$

$$= \sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^{\mathrm{T}} - N \cdot \bar{\boldsymbol{x}} \bar{\boldsymbol{x}}^{\mathrm{T}}$$

and the result again follows from the definition of a sample mean. □

Properties of key sample statistics, II (1/3)

Theorem 2

Properties of simple sample statistics (2). *Consider a random sample* $\{\boldsymbol{x}_i\}_{i=1}^N$ *drawn from a random vector* $\boldsymbol{x}$*, a transformation of this vector* $\boldsymbol{y} = \boldsymbol{g}(\boldsymbol{x})$*, and suppose that all the moments expressed in the mean vector* $\mathbb{E}[\boldsymbol{y}]$ *and in the variance-covariance matrix* $\mathbb{V}\text{ar}[\boldsymbol{y}]$ *are defined. The following two properties are true:*

a. $\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = N \cdot \mathbb{E}[\boldsymbol{y}_i]$;

b. $\mathbb{V}\text{ar}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = N \cdot \mathbb{V}\text{ar}[\boldsymbol{y}_i]$.

Proof.

To show **a.** simply observe that:

$$\mathbb{E}\left[\sum_{i=1}^N \boldsymbol{y}_i\right] = \sum_{i=1}^N \mathbb{E}[\boldsymbol{y}_i] = N \cdot \mathbb{E}[\boldsymbol{y}_i]$$

which follows from the linear properties of expectations and from the moments of $\boldsymbol{y}_i$ for $i = 1, \ldots, N$ being identical. **(Continues...)**

Properties of key sample statistics, II (2/3)

Theorem 2

Proof.

(Continued.) The demonstration of **b.** is as follows.

$$
\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\sum_{i=1}^{N} \boldsymbol{y}_i\right] &= \mathbb{E}\left[\left(\sum_{i=1}^{N} \boldsymbol{y}_i - \mathbb{E}\left[\sum_{i=1}^{N} \boldsymbol{y}_i\right]\right)\left(\sum_{i=1}^{N} \boldsymbol{y}_i - \mathbb{E}\left[\sum_{i=1}^{N} \boldsymbol{y}_i\right]\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\left(\sum_{i=1}^{N} \left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)\right)\left(\sum_{i=1}^{N} \left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)\right)^{\mathrm{T}}\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{N} \left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)\left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)^{\mathrm{T}}\right] \\
&= \sum_{i=1}^{N} \mathbb{E}\left[\left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)\left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)^{\mathrm{T}}\right] \\
&= N \cdot \mathbb{V}\mathrm{ar}\left[\boldsymbol{y}_i\right]
\end{aligned}
$$

(Continues. . .)

Properties of key sample statistics, II (3/3)

Theorem 2

Proof.

(Continued.) In the above derivation for **b.** the first line is just the definition of variance for $\sum_{i=1}^{N} \boldsymbol{y}_i$, the second line applies the linear properties of expectations while also rearranging terms, the third line rearranges terms again after observing that, for $i \neq j$:

$$\mathbb{E}\left[\left(\boldsymbol{y}_i - \mathbb{E}\left[\boldsymbol{y}_i\right]\right)\left(\boldsymbol{y}_j - \mathbb{E}\left[\boldsymbol{y}_j\right]\right)^{\mathrm{T}}\right] = \mathbf{0}$$

which follows from the independence of the realizations in the random sample, the fourth line is another application of the linear properties of expectations, while the fifth line again exploits the fact that all the realizations follow from identically distributed random variables. □

Note: independence (from samples being i.i.d.) is used to prove point **b.** but not point **a.** in the theorem.

Properties of key sample statistics, III (1/3)

Theorem 3

Properties of simple sample statistics (3). *Consider a random sample* $\{\boldsymbol{x}_i\}_{i=1}^N$ *drawn from a random vector* $\boldsymbol{x}$ *whose mean vector is* $\mathbb{E}[\boldsymbol{x}]$ *and whose variance-covariance matrix is* $\mathbb{V}\text{ar}[\boldsymbol{x}] < \infty$ *(finite). The following three properties are true:*

a. $\mathbb{E}[\bar{\boldsymbol{x}}] = \mathbb{E}[\boldsymbol{x}]$;

b. $\mathbb{V}\text{ar}[\bar{\boldsymbol{x}}] = \mathbb{V}\text{ar}[\boldsymbol{x}]/N$;

c. $\mathbb{E}[\boldsymbol{S}] = \mathbb{V}\text{ar}[\boldsymbol{x}]$.

Proof.

To show **a.** it is sufficient to apply Theorem 2, point **a.** for $\boldsymbol{y} = \boldsymbol{x}$:

$$\mathbb{E}[\bar{\boldsymbol{x}}] = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{x}_i] = \frac{1}{N}\cdot N\cdot\mathbb{E}[\boldsymbol{x}_i] = \mathbb{E}[\boldsymbol{x}]$$

(Continues...)

Properties of key sample statistics, III (2/3)

Theorem 3

Proof.
(Continued.) Point **b.** proceeds similarly.

$$\begin{aligned}
\mathbb{V}\mathrm{ar}\left[\bar{\boldsymbol{x}}\right] &= \mathbb{V}\mathrm{ar}\left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_i\right] \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i\right] \\
&= \frac{1}{N^2}\cdot N\cdot\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i\right] \\
&= \frac{\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]}{N}
\end{aligned}$$

(Continues...)

Properties of key sample statistics, III (3/3)

Theorem 3

Proof.

(Continues...) The proof of point **c.** is as follows:

$$\begin{aligned}
\mathbb{E}\left[\boldsymbol{S}\right] &= \mathbb{E}\left[\frac{1}{N-1}\left(\sum_{i=1}^{N}\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}} - N\cdot\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right)\right] \\
&= \frac{1}{N-1}\left(\sum_{i=1}^{N}\mathbb{E}\left[\boldsymbol{x}_i\boldsymbol{x}_i^{\mathrm{T}}\right] - N\cdot\mathbb{E}\left[\bar{\boldsymbol{x}}\bar{\boldsymbol{x}}^{\mathrm{T}}\right]\right) \\
&= \frac{1}{N-1}\left(N\cdot\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i\right] - N\cdot\mathbb{V}\mathrm{ar}\left[\bar{\boldsymbol{x}}\right]\right) \\
&= \frac{N}{N-1}\left(1-\frac{1}{N}\right)\mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right] \\
&= \mathbb{V}\mathrm{ar}\left[\boldsymbol{x}\right]
\end{aligned}$$

the third line follows after adding and subtracting $N\cdot\mathbb{E}\left[\boldsymbol{x}\right]\mathbb{E}\left[\boldsymbol{x}\right]^{\mathrm{T}}$. □

Normal sampling

- While performing statistical estimation and inference, it is often useful to know the **exact sampling distribution** of selected statistics like $\bar{\boldsymbol{x}}$ and $\boldsymbol{S}$.
- This is usually possible only in selected cases, like the one where the sample is drawn from a **normal distribution**.
- Let sample $\{x_i\}_{i=1}^N$ be drawn from $X \sim \mathcal{N}(\mu, \sigma^2)$; then:

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

  follows from the property of (independent) normal r.v.s.
- An equivalent, convenient formulation also follows.

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

The *t*-statistic

- The "standardized" statistic $\sqrt{N}\left(\bar{X}-\mu\right)/\sigma$ is appealing for the sake of testing hypotheses about μ.
- However, it has a shortcoming: usually, σ is **unknown**.
- Intuitively, one could replace σ with the sample standard deviation S. This gives rise to the following statistic.

Definition 9

The *t*-statistic. Given a univariate sample $\{x_i\}_{i=1}^{N}$ of size N drawn from a sequence of random variables $X_1, \ldots, X_N$, a *t-statistic* is defined as the following quantity:

$$\mathrm{t} = \sqrt{N}\,\frac{\bar{X}-\mu}{S}$$

where $\bar{X}$ is the sample mean whose expectation is $\mu = \mathbb{E}\left[\bar{X}\right]$, and S is the sample standard deviation.

Properties of normal sampling (1/6)

The exact sampling distribution of the t-statistic is well-known, but deriving it requires a few steps.

Theorem 4

Sampling from the Normal Distribution. *Consider a random sample* $\{x_i\}_{i=1}^{N}$ *drawn from a random variable following the normal distribution* $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, *and the random variables corresponding to the two sample statistics* $\bar{X}$ *and* S^2. *The following three properties are true:*

a. $\bar{X}$ *and* S^2 *are independent;*

b. $\bar{X} \sim \mathcal{N}\left(\mu, \sigma^2/N\right)$;

c. $(N-1)\, S^2/\sigma^2 \sim \chi^2_{N-1}$.

Proof.

Point **b.** is straightforward, point **c.** is quite easy to show, but point **a.** requires some more effort. **(Continues...)**

Properties of normal sampling (2/6)

Theorem 4

Proof.

(Continued.) Start by observing that the sample variance can be expressed in terms of only $N-1$ of the original random variables:

$$\begin{aligned} S^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(X_i-\bar{X}\right)^2 \\ &= \frac{1}{N-1}\left[\left(X_1-\bar{X}\right)^2+\sum_{i=2}^{N}\left(X_i-\bar{X}\right)^2\right] \\ &= \frac{1}{N-1}\left[\left(\sum_{i=2}^{N}\left(X_i-\bar{X}\right)\right)^2+\sum_{i=2}^{N}\left(X_i-\bar{X}\right)^2\right] \end{aligned}$$

where the last line follows from $\sum_{i=1}^{N}\left(X_i-\bar{X}\right)=0$. Hence, proving that the sample mean is independent of the sample variance amounts to show that it is independent of $N-1$ out of N normally distributed random variables, say $X_2-\bar{X},\ldots,X_N-\bar{X}$. **(Continues...)**

Properties of normal sampling (3/6)

Theorem 4

Proof.

(Continued.) Work with the standardization $Z_i = (X_i - \mu)/\sigma$ for $i = 1, \ldots, N$, and let $\boldsymbol{z} = (Z_i, \ldots, Z_N)$. Define the following random vector $\widetilde{\boldsymbol{z}}$ of length N as a function of $\boldsymbol{z}$.

$$\widetilde{\boldsymbol{z}} = \begin{pmatrix} \bar{Z} \\ \widetilde{Z}_2 \\ \vdots \\ \widetilde{Z}_N \end{pmatrix} = \begin{pmatrix} \bar{Z} \\ Z_2 - \bar{Z} \\ \vdots \\ Z_N - \bar{Z} \end{pmatrix} = \begin{bmatrix} N^{-1} & N^{-1} & \ldots & N^{-1} \\ -N^{-1} & 1 - N^{-1} & \ldots & -N^{-1} \\ \vdots & \vdots & \ddots & \vdots \\ -N^{-1} & -N^{-1} & \ldots & 1 - N^{-1} \end{bmatrix} \boldsymbol{z}$$

One shall show that $\widetilde{\boldsymbol{z}}$ is composed of independent random variables:

- this would show that $\bar{Z}$ is independent of $Z_2 - \bar{Z}, \ldots, Z_N - \bar{Z}$;
- so, $\bar{X}$ would be independent of $X_2 - \bar{X}, \ldots, X_N - \bar{X}$ (and of S^2).

The above transformation is *linear* and its Jacobian has determinant $1/N$, thus it is invertible. By the properties of linear transformations, its *inverse* has a Jacobian with determinant N. **(Continues...)**

Properties of normal sampling (4/6)

Theorem 4

Proof.

(Continued.) The joint p.d.f. of $\widetilde{\boldsymbol{z}}$ obtains directly from that of $\boldsymbol{z}$:

$$
\begin{aligned}
f_{\widetilde{\boldsymbol{z}}}\left(\bar{z},\widetilde{z}_2,\ldots,\widetilde{z}_N\right) &= \frac{N}{\sqrt{\left(2\pi\right)^N}}\exp\left(-\frac{1}{2}\left(\bar{z}-\sum_{i=2}^{N}\widetilde{z}_i\right)^2-\frac{1}{2}\sum_{i=2}^{N}\left(\bar{z}+\widetilde{z}_i\right)^2\right)\\
&= \sqrt{\frac{N}{2\pi}}\exp\left(-\frac{N\bar{z}^2}{2}\right)\times\\
&\quad\times\sqrt{\frac{N}{\left(2\pi\right)^{N-1}}}\exp\left(-\frac{1}{2}\left(\sum_{i=2}^{N}\widetilde{z}_i\right)^2-\frac{1}{2}\sum_{i=2}^{N}\widetilde{z}_i^2\right)\\
&= f_{\bar{Z}}\left(\bar{z}\right)\cdot f_{\widetilde{\boldsymbol{z}}_{-1}}\left(\widetilde{z}_2,\ldots,\widetilde{z}_N\right)
\end{aligned}
$$

and it can be clearly decomposed into the product of two components: the p.d.f. of $\bar{Z}$ and that of all the other elements of $\widetilde{\boldsymbol{z}}$. Therefore, $\bar{Z}$ is independent of $Z_2-\bar{Z},\ldots,Z_N-\bar{Z}$: **a.** is proved. **(Continues...)**

Properties of normal sampling (5/6)

Theorem 4

Proof.

(Continued.) Moving to the other points, **b.** as argued is obvious, while to demonstrate point **c.** it is easiest to proceed as follows.

$$
\begin{aligned}
(N-1)\,\frac{S^2}{\sigma^2} &= \sum_{i=1}^{N} \frac{\left(X_i - \bar{X}\right)^2}{\sigma^2} \\
&= \sum_{i=1}^{N} \frac{\left(X_i - \mu + \mu - \bar{X}\right)^2}{\sigma^2} \\
&= \sum_{i=1}^{N} \frac{\left(X_i - \mu\right)^2}{\sigma^2} - \frac{N\left(\bar{X} - \mu\right)^2}{\sigma^2} - 2\left(\bar{X} - \mu\right) \sum_{i=1}^{N} \frac{X_i - \mu}{\sigma^2} \\
&= \sum_{i=1}^{N} \left(\frac{X_i - \mu}{\sigma}\right)^2 - \left(\sqrt{N}\,\frac{\bar{X} - \mu}{\sigma}\right)^2
\end{aligned}
$$

(Continues...)

Properties of normal sampling (6/6)

Theorem 4

Proof.

(Continued.) The statistic $(N-1)\,S^2/\sigma^2$ is shown to be the sum of the *squares* of N independent random variables – all of which follow the standard normal distribution, *minus* the square of another random variable that follows the standard normal distribution. By the result in **a.** the latter is independent of the former. Note that, using m.g.f.s:

$$M_{\bar{\bar{Z}}^2}(t)\, M_{(N-1)\frac{S^2}{\sigma^2}}(t) = \prod_{i=1}^{N} M_{Z_i^2}(t)$$

where $\bar{\bar{Z}} \equiv \sqrt{N}\left(\bar{X}-\mu\right)/\sigma$ and $Z_i \equiv (X_i-\mu)/\sigma$, or equivalently:

$$M_{(N-1)\frac{S^2}{\sigma^2}}(t) = \frac{1}{M_{\bar{\bar{Z}}^2}(t)} \prod_{i=1}^{N} M_{Z_i^2}(t) = (1-2t)^{-\frac{1}{2}(N-1)}$$

following since $\bar{\bar{Z}}^2, Z_i^2 \sim \chi_1^2$. Therefore, $(N-1)\,S^2/\sigma^2 \sim \chi_{N-1}^2$: point **c.** is proved too. □

A *t*-distribution for the *t*-statistic

Owning these results, it is possible to return to the *t*-statistic and derive its distribution. Note that:

$$\mathrm{t} = \sqrt{N}\,\frac{\bar{X}-\mu}{S} = \frac{\sqrt{N}\,\dfrac{\bar{X}-\mu}{\sigma}}{\sqrt{(N-1)\,\dfrac{S^2}{\sigma^2}\,\dfrac{1}{\sqrt{N-1}}}} \sim \mathcal{T}_{N-1}$$

is the ratio between two independent random variables:

- the numerator follows the standard normal distribution,
- while the denominator equals the square root of a random variable following the chi-squared distribution with $N-1$ degrees of freedom, *divided* by the square root of $N-1$.

Hence, by Observation 2 from Lecture 3, a *t*-statistic follows the Student's *t*-distribution with $N-1$ degrees of freedom.

The *F*-statistic

Knowing the distribution of the t-statistic helps conduct tests about μ in a normal random sample, but what about σ^2?

Definition 10

Normal variance ratio. Consider two univariate random samples $\{x_i\}_{i=1}^{N_X}$ and $\{y_i\}_{i=1}^{N_Y}$ of sizes N_X and N_Y respectively, each drawn from two *independent* sequences of random variables $(X_1, \ldots, X_{N_X})$ and $(Y_1, \ldots, Y_{N_Y})$ whose distributions are as follows.

$$X_i \sim \mathcal{N}\left(\mu_X, \sigma_X^2\right) \text{ for } i = 1, \ldots, N_X$$
$$Y_j \sim \mathcal{N}\left(\mu_Y, \sigma_Y^2\right) \text{ for } j = 1, \ldots, N_Y$$

The normal variance ratio is defined as the following F-statistic:

$$\mathrm{F} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2}$$

where S_X & S_Y are the sample variances of the two random samples.

An *F*-distribution for the *F*-statistic

- The F-statistic is used to test whether any two populations have identical/similar variance.

- To this end, one shall know the exact sampling distribution of the F-statistic: again, random ratios come to rescue.

- Both the numerator and denominator of F, if multiplied by $N_X - 1$ and $N_Y - 1$ respectively, follow – by Theorem 4 – a chi-squared distribution with those given numbers as their respective degrees of freedom.

- Therefore, by Observation 3 from Lecture 3:

$$\mathrm{F} = \frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \sim \mathcal{F}_{N_X-1,N_Y-1}$$

  a result that can be exploited in statistical inference.

Multivariate normal sampling

- The analysis of normal sampling so far concerns univariate samples. These ideas also extend to a multivariate setting.

- Specifically, let the random sample $\{\mathbf{x}_i\}_{i=1}^N$ be drawn from some multivariate normal distribution $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

- By the properties of i.i.d. samples and of the multivariate normal distribution, the following holds.

$$\bar{\boldsymbol{x}} \sim \mathcal{N}\left(\boldsymbol{\mu}, \frac{\boldsymbol{\Sigma}}{N}\right)$$

- In statistical inference and in testing hypotheses, interest usually falls on the vector of means $\boldsymbol{\mu}$.

- A *scalar* statistic that summarizes while standardizing all sample means appears useful here.

The u-statistic.

Definition 11

The u-statistic. Given a multivariate sample $\{\mathbf{x}_i\}_{i=1}^{N}$ of size N drawn from a sequence of random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, a *u-statistic* is defined as the following quantity:

$$\begin{aligned} \mathrm{u} &= N\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \\ &= N \sum_{k=1}^{K} \sum_{\ell=1}^{K} \sigma_{k\ell}^{*-1} \left(\bar{X}_k - \mu_k\right)\left(\bar{X}_\ell - \mu_\ell\right) \end{aligned}$$

where here $\bar{\boldsymbol{x}}$ is the sample mean whose expectation and variance are $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}_i\right]$, and $\boldsymbol{\Sigma}/N = \mathbb{V}\mathrm{ar}\left[\boldsymbol{x}_i\right]$ respectively, and where $\sigma_{k\ell}^{*-1}$ is $k\ell$-th element of $\boldsymbol{\Sigma}^{-1}$.

To better interpret the u-statistic, one might analyze its expression as a quadratic form in the second line of the definition: the statistic is a second degree polynomial of the K deviations of all univariate sample means from their respective mean parameters, *normalized* through the population variance-covariance.

The distribution of the u-statistic (1/2)

One more time, using this statistic for inference purposes requires the derivation of its exact distribution.

Theorem 5

Sampling from the Multivariate Normal Distribution. *Let a random sample* $\{\mathbf{x}_i\}_{i=1}^N$ *be drawn from some* K*-dimensional random vector following the multivariate normal distribution,* $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$*. In this environment, the u-statistic follows the chi-squared distribution with* K *degrees of freedom.*

$$\mathrm{u} = N(\bar{\boldsymbol{x}} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\bar{\boldsymbol{x}} - \boldsymbol{\mu}) \sim \chi^2_K$$

Proof.

As usual, the result requires to find the m.g.f. of the random variable of interest: here, the u-statistic. **(Continues...)**

The distribution of the u-statistic (2/2)

Theorem 5

Proof.

(Continued.) This requires some linear algebra.

$$
\begin{aligned}
M_{\mathrm{u}}(t) &= \int_{\mathbb{R}^K} \exp\left(N\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)t\right) f_{\bar{\boldsymbol{x}}}\left(\bar{\mathbf{x}};\boldsymbol{\mu},\frac{\boldsymbol{\Sigma}}{N}\right)\mathrm{d}\bar{\mathbf{x}} \\
&= \int_{\mathbb{R}^K} \sqrt{\frac{1}{(2\pi)^K}\frac{N^K}{|\boldsymbol{\Sigma}|}} \exp\left(-\frac{N}{2}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}}\left[\left(1-2t\right)\boldsymbol{\Sigma}^{-1}\right]\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)t\right)\mathrm{d}\bar{\mathbf{x}} \\
&= \sqrt{\frac{1}{(1-2t)^K}} \times \int_{\mathbb{R}^K} \sqrt{\frac{1}{(2\pi)^K}\frac{\left[\left(1-2t\right)N\right]^K}{|\boldsymbol{\Sigma}|}} \times \\
&\qquad \times \exp\left(-\frac{N}{2}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}}\left[\left(1-2t\right)\boldsymbol{\Sigma}^{-1}\right]\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)t\right)\mathrm{d}\bar{\mathbf{x}} \\
&= (1-2t)^{-\frac{K}{2}}
\end{aligned}
$$

Note: the integral in the third line disappears since it is the p.d.f. of a multivariate normal distribution with mean $\boldsymbol{\mu}$ and variance-covariance $\boldsymbol{\Sigma}/\left(1-2t\right)N$. By recognizing the m.g.f. it follows that $\mathrm{u} \sim \chi^2_K$. $\square$

Hotelling's *t*-squared statistic

Problems reiterate! Like in the univariate case, if $\boldsymbol{\Sigma}$ is unknown the u-statistic is useless. What if $\boldsymbol{\Sigma}$ is replaced by its sample analog $\boldsymbol{S}$?

Definition 12

Hotelling's "*t*-squared" statistic. Given some multivariate sample $\{\mathbf{x}_i\}_{i=1}^{N}$ of size N drawn from a sequence of random vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$, Hotelling's *t-squared statistic* is defined as the random variable:

$$\begin{aligned} \mathrm{t}^2 &= N\left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \\ &= N \sum_{k=1}^{K} \sum_{\ell=1}^{K} S_{k\ell}^{*-1} \left(\bar{X}_k - \mu_k\right) \left(\bar{X}_\ell - \mu_\ell\right) \end{aligned}$$

where $\bar{\boldsymbol{x}}$ is the sample mean whose expectation is $\boldsymbol{\mu} = \mathbb{E}\left[\boldsymbol{x}_i\right]$, $\boldsymbol{S}$ is the sample variance-covariance, and $S_{k\ell}^{*-1}$ is $k\ell$-th element of $\boldsymbol{S}^{-1}$.

One can prove that a *rescaled* version of t^2 follows the F-distribution with paired degrees of freedom K and $N - K$.

$$\frac{N-K}{K\left(N-1\right)} \mathrm{t}^2 = \frac{N\left(N-K\right)}{K\left(N-1\right)} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right)^{\mathrm{T}} \boldsymbol{S}^{-1} \left(\bar{\boldsymbol{x}} - \boldsymbol{\mu}\right) \sim \mathcal{F}_{K,N-K}$$

Order statistics

It is often useful to study the values that the realizations of a random variable take at a given position in the *order* of realizations (smallest value, highest value, *et cetera*).

Definition 13

Order statistics. Consider a sample $\{x_i\}_{i=1}^N$ of realizations obtained from univariate random variables, $\{X_i\}_{i=1}^N$. Suppose that these values are placed in *ascending order*, where subscripts surrounded by parentheses denote one observation's position in the order:

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(N)}$$

therefore, $x_{(1)} = \min\{x_i\}_{i=1}^N$ and $x_{(N)} = \max\{x_i\}_{i=1}^N$. The j-th *order statistic* is the random variable – denoted as $X_{(j)}$ – that generates the j-th realization in the above sequence, that is $x_{(j)}$.

Every univariate sample has N associated order statistics that need to satisfy the following property.

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(N)}$$

Minima, maxima, ranges and medians

Definition 14

Sample Minimum. The sample *minimum* is the first order statistic, $X_{(1)}$.

Definition 15

Sample Maximum. The sample *maximum* is the N-th order statistic, $X_{(N)}$.

Definition 16

Sample Range. The sample *range* R is the difference between the sample maximum and the sample minimum: $R = X_{(N)} - X_{(1)}$.

Definition 17

Sample Median. The sample *median* M is a function of a sample's most central order statistics.

$$M = \begin{cases} X_{\left(\frac{N+1}{2}\right)} & \text{if } N \text{ is odd} \\ \frac{1}{2}\left(X_{\left(\frac{N}{2}\right)} + X_{\left(\frac{N}{2}+1\right)}\right) & \text{if } N \text{ is even} \end{cases}$$

Which sampling distribution for order statistics?

- One may want to provide a distribution for order statistics (e.g. to model the probabilities for minima and maxima).

- If the sample is random, the c.d.f. of the j-th order statistic can be expressed in terms of the following joint probability.

$$F_{X_{(j)}}(x) = \mathbb{P}\Big(X_{(1)} \leq x \cap \cdots \cap X_{(j)} \leq x \cap X_{(j+1)} > x \cap \cdots \cap X_{(N)} > x\Big)$$

- This allows to provide formulae for the p.d.f.s and c.d.f.s of order statistics, although these can be hard to use.

- There are some relevant cases: order statistics for uniform distributions, minima or maxima for selected distributions.

Sampling distribution of order statistics (1/5)

Theorem 6

Sampling distribution of order statistics in random samples. *In a univariate random sample, the c.d.f. of the j-th order statistic is based on the binomial distribution:*

$$F_{X_{(j)}}(x) = \sum_{k=j}^{N} \binom{N}{k} \left[F_X(x)\right]^k \left[1 - F_X(x)\right]^{N-k}$$

where $F_X(x)$ is the c.d.f. of the random variable X that generates the sample. Two particular cases are the c.d.f.s of the minimum and the maximum, which are as follows.

$$F_{X_{(1)}}(x) = 1 - \left[1 - F_X(x)\right]^N$$
$$F_{X_{(N)}}(x) = \left[F_X(x)\right]^N$$

Proof.

(Continues...)

Sampling distribution of order statistics (2/5)

Theorem 6

Proof.

(Continued.) For precisely j realizations to be less or equal than x and the other $N-j$ to be larger than x, the event defined as $X_i \leq x$ must occur j times, whereas the complementary event $X_i > x$ must occur $N-j$ times. If the sample is random (i.i.d.), these two events occur with probabilities that are constant across all realizations.

$$\mathbb{P}\left(X_i \leq x\right) = F_X\left(x\right)$$
$$\mathbb{P}\left(X_i > x\right) = 1 - F_X\left(x\right)$$

As the sample is random, all joint combinations of said events can be expressed as the appropriate product of those probabilities. Obviously, the joint events follow a binomial distribution, where the count of all potential combinations with j 'successes' $(X_i \leq x)$ and $N-j$ 'failures' $(X_i > x)$ is provided by the binomial coefficient. The distributions for the minimum and the maximum are special cases of this result. $\square$

Sampling distribution of order statistics (3/5)

If the sample is drawn from a continuous distribution, one can also derive the p.d.f. of an order statistic of interest.

Corollary

(Theorem 6.) *If X is a continuous distribution with p.d.f. $f_X(x)$, the p.d.f. of the j-th order statistic is the following.*

$$f_{X_{(j)}}(x) = \frac{N!}{(j-1)!\,(N-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{N-j}$$

Proof.

The result obtains by taking the first derivative of $F_{X_{(j)}}(x)$ and some manipulation. The initial differentiation is shown in the next slide; it applies the chain rule while subsequently, the third line in the display is obtained by isolating the term corresponding with $k = j$ in the sum that results from taking the derivative. Then, one needs to show that the two elements that are left out cancel out against one another.

(Continues...)

Sampling distribution of order statistics (4/5)

Proof.

(Continued.) The initial operations are as follows.

$$
\begin{aligned}
f_{X_{(j)}}(x) &= \frac{dF_{X_{(j)}}(x)}{dx} \\
&= \sum_{k=j}^{N} \binom{N}{k} \Big(k \left[F_X(x)\right]^{k-1} \left[1 - F_X(x)\right]^{N-k} f_X(x) - \\
&\quad - (N-k) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x) \Big) \\
&= \frac{N!}{(j-1)!\,(N-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{N-j} + \\
&\quad + \sum_{k=j+1}^{N} \binom{N}{k} k \left[F_X(x)\right]^{k-1} \left[1 - F_X(x)\right]^{N-k} f_X(x) - \\
&\quad - \sum_{k=j}^{N} \binom{N}{k} (N-k) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x)
\end{aligned}
$$

(Continues...)

Sampling distribution of order statistics (5/5)

Proof.

(Continued.) Re-index the summation of the second term above and ntoe that in the summation of the third term, the element for $N = k$ is zero. Thus, the p.d.f. can be rewritten as follows.

$$f_{X_{(j)}}(x) = \frac{N!}{(j-1)!\,(N-j)!} f_X(x) \left[F_X(x)\right]^{j-1} \left[1 - F_X(x)\right]^{N-j} +$$
$$+ \sum_{k=j}^{N-1} \binom{N}{k+1} (k+1) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x) -$$
$$- \sum_{k=j}^{N-1} \binom{N}{k} (N-k) \left[F_X(x)\right]^{k} \left[1 - F_X(x)\right]^{N-k-1} f_X(x)$$

Since, by simple manipulation of factorials, it is:

$$\binom{N}{k+1} (k+1) = \frac{N!}{k!\,(N-k-1)!} = \binom{N}{k} (N-k)$$

it follows that the two terms in question cancel out. □

Order statistics for the uniform distribution

These formulae can seldom be linked to any known distributions. One notable case is the following.

Observation 1

Consider a random sample obtained from the standard continuous uniform distribution, $X \sim \mathcal{U}(0,1)$. The j-th order statistic is such that $X_{(j)} \sim \text{Beta}(j, N-j+1)$.

Proof.

Since $F_X(x) = x$ and $f_X(x) = 1$ for $x \in (0,1)$, while $F_X(x) = x$ and $f_X(x) = 0$ otherwise, the density function of $X_{(j)}$ is, for $x \in (0,1)$:

$$\begin{aligned} f_{X_{(j)}}(x) &= \frac{N!}{(j-1)!\,(N-j)!}\, x^{j-1}\,(1-x)^{N-j} \\ &= \frac{\Gamma(N+1)}{\Gamma(j)\,\Gamma(N-j+1)}\, x^{j-1}\,(1-x)^{(N-j+1)-1} \\ &= \text{B}(j, N-j+1) \cdot x^{j-1}\,(1-x)^{(N-j+1)-1} \end{aligned}$$

the result is the p.d.f. of the postulated Beta distribution. □

Min-max stability

- The result about this uniform distribution is useful: it applies to *percentiles* p drawn from any distribution, since $p \sim \mathcal{U}(0, 1)$.

- Other results are specific to the minima and maxima: sometimes these two *extreme* order statistics follow yet another distribution from the same family whence the sample is drawn.

- This deserves a proper definition.

Definition 18

Extreme order statistics (minimum-maximum) stability. Consider a random sample drawn from some known distribution. If the sample minimum (maximum) follows another distribution of the same family, that distribution is said to be *min-stable* (*max-stable*).

- Min-/max-stable distributions include the exponential as well as the GEV distributions.

Min-stability of the exponential distribution

Observation 2

Consider a random sample drawn from the exponential distribution with parameter λ, $X \sim \text{Exp}(\lambda)$. In this case, the first order statistic (the minimum) is such that $X_{(1)} \sim \text{Exp}\left(N^{-1}\lambda\right)$.

Proof.

By applying the formula for the distribution of the minimum:

$$\begin{aligned} F_{X_{(1)}}(x;\lambda,N) &= 1-\left[1-F_X(x;\lambda)\right]^N \\ &= 1-\left[\exp\left(-\frac{1}{\lambda}x\right)\right]^N \\ &= 1-\exp\left(-\frac{N}{\lambda}x\right) \end{aligned}$$

the postulated c.d.f. obtains directly. □

Max-stability of the Gumbel distribution

Observation 3

Consider a random sample drawn from the Type I GEV (Gumbel) distribution with parameters μ and σ, $X \sim \mathrm{EV1}(\mu, \sigma)$. The top order statistic (the maximum) is such that $X_{(N)} \sim \mathrm{EV1}(\mu - \sigma \log(N), \sigma)$.

Proof.

By applying the formula for the distribution of the maximum:

$$\begin{aligned} F_{X_{(N)}}(x; \mu, \sigma, N) &= \left[F_X(x; \mu, \sigma)\right]^N \\ &= \exp\left(-\exp\left(\frac{x-\mu}{\sigma}\right)\right)^N \\ &= \exp\left(-N\exp\left(\frac{x-\mu}{\sigma}\right)\right) \\ &= \exp\left(-\exp\left(\frac{x-\mu+\sigma\log(N)}{\sigma}\right)\right) \end{aligned}$$

one obtains the Gumbel c.d.f. that was argued. □

Max-stability in other GEV distributions

Observation 4

Consider a random sample drawn from the Type II GEV (Fréchet) distribution with parameters α, μ, and σ, $Y \sim \text{EV2}(\alpha, \mu, \sigma)$. The top order statistic (the maximum) is such that $Y_{(N)} \sim \text{EV2}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$. The result is identical in the Type III GEV (reverse Weibull) case: if $Y \sim \text{EV3}(\alpha, \mu, \sigma)$, it is $Y_{(N)} \sim \text{EV3}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$.

Proof.

Here, applying the formula for the distribution of the maximum:

$$\begin{aligned} F_{Y_{(N)}}(y; \alpha, \mu, \sigma, N) &= \left[F_Y(y; \alpha, \mu, \sigma)\right]^N \\ &= \exp\left(-\left(\frac{y-\mu}{\sigma}\right)^{-\alpha}\right)^N \\ &= \exp\left(-\left[N^{-\frac{1}{\alpha}}\left(\frac{y-\mu}{\sigma}\right)\right]^{-\alpha}\right) \end{aligned}$$

allows to show both the Fréchet and the reverse Weibull results. □

Min-stability of the Weibull distribution

Observation 5

Consider a random sample that is drawn from the traditional Weibull distribution with parameters α, μ, and σ, $W \sim \text{Weibull}\,(\alpha, \mu, \sigma)$. The sample minimum is such that $W_{(1)} \sim \text{Weibull}\left(\alpha, \mu, \sigma N^{1/\alpha}\right)$.

Proof.

Things proceed similarly as in the Fréchet and reverse Weibull cases.

$$\begin{aligned} F_{W_{(1)}}\left(w; \alpha, \mu, \sigma, N\right) &= 1 - \left[1 - F_W\left(w; \alpha, \mu, \sigma\right)\right]^N \\ &= 1 - \exp\left(-\left(\frac{w-\mu}{\sigma}\right)^{-\alpha}\right)^N \\ &= 1 - \exp\left(-\left[N^{-\frac{1}{\alpha}}\left(\frac{w-\mu}{\sigma}\right)\right]^{-\alpha}\right) \end{aligned}$$

Note that here, the formula for the minimum is applied instead. $\square$

Sufficient Statistics

The final kind of statistic that is introduced in this lecture serve as a first step for the study of estimation (Lecture 5).

Definition 19

Sufficient statistics. Consider a given sample generated by a list of random vectors $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$. Suppose that the joint distribution of the sample depends, among the others, on some parameter θ; write the associated p.m.f. or p.d.f. as $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)$. A statistic $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is said to be *sufficient* if the joint distribution of the sample, conditional on it, does not depend on θ:

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N|T}(\mathbf{x}_1, \ldots, \mathbf{x}_N | T(\mathbf{x}_1, \ldots, \mathbf{x}_N)) = \frac{f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)}{q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)}$$

where $q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$ is the p.m.f. or the p.d.f. of the sufficient statistic in question.

This can also be expressed by saying that the joint conditional density is constant as a function of θ.

The Sufficiency Principle

- The intuition behind sufficient statistics is that they "exhaust" all the information about θ that is contained in a sample.
- This aids estimation and inference in various ways.
- The role of sufficient statistics in inference is summarized by the following **statistical principle**. This is a postulate (an axiom) of statistical analysis.

Statistical Principle 1. Sufficiency. If $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a sufficient statistic for a parameter θ, any evaluation about the latter should depend solely on the sufficient statistic or a function thereof. That is, if $(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ and $(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ are two, possibly different sample realizations such that

$$T(\mathbf{x}_1, \ldots, \mathbf{x}_N) = T(\mathbf{y}_1, \ldots, \mathbf{y}_N)$$

all statistical evaluations about θ should be identical regardless of the exact observed values in either realization.

Example: a Bernoulli sufficient statistic

- Consider a random sample drawn from $X \sim \mathrm{Be}(p)$.
- The count of "successes" is a sufficient statistic for p.

$$T = T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$$

Its *realization* is $t = T(x_1, \ldots, x_N) = \sum_{i=1}^{N} x_i$

- Apply the definition, observing that $T \sim \mathrm{BN}(p, N)$.

$$\begin{aligned}
\frac{f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; p)}{q_T(t; p, N)} &= \frac{\prod_{i=1}^{N} p^{x_i} (1-p)^{1-x_i}}{\binom{N}{t} p^t (1-p)^{N-t}} \\
&= \frac{p^t (1-p)^{N-t}}{\binom{N}{t} p^t (1-p)^{N-t}} \\
&= \frac{t!\,(N-t)!}{N!}
\end{aligned}$$

Example: $\bar{X}$ suffices for the normal's μ (1/2)

- Consider a random sample drawn from $X \sim \mathcal{N}(\mu, \sigma^2)$.
- The sample mean $\bar{X}$ is a sufficient statistic for μ.
- Recall that $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$, and write the realization of $\bar{X}$ follows.

$$\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$$

- In the application that follows next, the following property is applied (as point **c.** of Theorem 4).

$$\sum_{i=1}^{N} (x_i - \mu)^2 = \sum_{i=1}^{N} (x_i - \bar{x})^2 + N(\bar{x} - \mu)^2 - 2(\bar{x} - \mu) \underbrace{\sum_{i=1}^{N} (x_i - \bar{x})}_{=0}$$

Example: $\bar{X}$ suffices for the normal's μ (2/2)

The derivation follows suit; note how the mentioned property is applied in the second line.

$$\begin{aligned}
\frac{f_{X_1,\ldots,X_N}\left(x_1,\ldots,x_N;\mu,\sigma^2\right)}{q_{\bar{X}}\left(\bar{x};\mu,\sigma^2/N\right)} &= \frac{\prod_{i=1}^{N}\sqrt{\left(2\pi\sigma^2\right)^{-1}}\cdot\exp\left(-\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)}{\sqrt{\left(2\pi\sigma^2\right)^{-1}N}\cdot\exp\left(-\frac{N\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right)} \\
&= \frac{\sqrt{\left(2\pi\sigma^2\right)^{-N}}\cdot\exp\left(-\sum_{i=1}^{N}\frac{\left(x_i-\mu\right)^2}{2\sigma^2}\right)}{\sqrt{\left(2\pi\sigma^2\right)^{-1}N}\cdot\exp\left(-\frac{N\left(\bar{x}-\mu\right)^2}{2\sigma^2}\right)} \\
&= \frac{\exp\left(-\sum_{i=1}^{N}\frac{\left(x_i-\bar{x}\right)^2}{2\sigma^2}\right)}{\sqrt{\left(2\pi\sigma^2\right)^{N-1}N}}
\end{aligned}$$

Example: a sufficient order statistic

- Consider a random sample drawn from $X \sim \mathcal{U}(0, \theta)$.
- The maximum $X_{(N)}$ is a sufficient statistic for parameter θ.
- The realization of $X_{(N)}$ is $x_{(N)} = \max\{x_1, \ldots, x_N\}$, and:

$$q_{X_{(N)}}\left(x_{(N)}; \theta\right) = \frac{d}{dx_{(N)}}\left[\left(\frac{x_{(N)}}{\theta}\right)^N \cdot \mathbb{1}\left[x_{(N)} \in (0, \theta)\right]\right]$$
$$= \frac{N x_{(N)}^{N-1}}{\theta^N} \cdot \mathbb{1}\left[x_{(N)} \in (0, \theta)\right]$$

- . . . hence, applying the definition here gives:

$$\frac{f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \theta)}{q_{X_{(N)}}\left(x_{(N)}; \theta\right)} = \frac{1}{N x_{(N)}^{N-1}} \cdot \mathbb{1}\left[x_{(N)} \in (0, \theta)\right]$$

since here $f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \theta) = [f_X(x; \theta)]^N = \theta^{-N}$.

The factorization theorem (1/7)

The following important result helps identify sufficient statistics.

Theorem 7

Fisher-Neyman's Factorization Theorem. *Consider a sample generated by a sequence of random vectors* $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$*, whose joint distribution has a p.m.f. or a p.d.f.* $f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)$ *that also depends on some parameter* θ*. Then, a statistic* $T = T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ *is sufficient for* θ *if and only if it is possible to identify two functions* $g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$ *and* $h(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *such that the following holds.*

$$f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta) = g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta) \cdot h(\mathbf{x}_1, \ldots, \mathbf{x}_N)$$

Observe that function $g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)$ *depends on* θ*, but function* $h(\mathbf{x}_1, \ldots, \mathbf{x}_N)$ *does not.*

Proof.

This proof is complex, and is only fully developed in the discrete case; the continuous case is only *sketched.* **(Continues...)**

The factorization theorem (2/7)

Theorem 7

Proof.

(Continued.) Start from the discrete case and the "necessity" part: if the factorization exists, then T is sufficient for θ. Write the p.m.f. of T as $q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)$. Furthermore, define the set of vectors spanning the same space as $(\mathbf{x}_1,\ldots,\mathbf{x}_N)$ and that result in the same values for T, as follows.

$$\mathbb{A}_T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\equiv\left\{\mathbf{y}_1,\ldots,\mathbf{y}_N : T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)=T\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right)\right\}$$

By the property of probability functions, the following holds.

$$\begin{aligned} q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) &= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{y}_1,\ldots,\mathbf{y}_N;\theta\right) \\ &= \sum_{\mathbf{y}_1,\ldots,\mathbf{y}_N\in\mathbb{A}_T} g\left(T\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right);\theta\right) h\left(\mathbf{y}_1,\ldots,\mathbf{y}_N\right) \end{aligned}$$

(Continues...)

The factorization theorem (3/7)

Theorem 7

Proof.

(Continued.) Since $T(\mathbf{y}_1, \ldots, \mathbf{y}_N)$ is constant in $\mathbb{A}_T(\mathbf{x}_1, \ldots, \mathbf{x}_N)$, it is:

$$q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta) = g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta) \sum_{\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{A}_T} h(\mathbf{y}_1, \ldots, \mathbf{y}_N)$$

where in both cases $\mathbb{A}_T$ is shorthand notation for $\mathbb{A}_T(\mathbf{x}_1, \ldots, \mathbf{x}_N)$. It then follows that:

$$\begin{aligned} \frac{f_{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N}(\mathbf{x}_1, \ldots, \mathbf{x}_N; \theta)}{q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)} &= \frac{g(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta) \cdot h(\mathbf{x}_1, \ldots, \mathbf{x}_N)}{q_T(T(\mathbf{x}_1, \ldots, \mathbf{x}_N); \theta)} \\ &= \frac{h(\mathbf{x}_1, \ldots, \mathbf{x}_N)}{\sum_{\mathbf{y}_1, \ldots, \mathbf{y}_N \in \mathbb{A}_T} h(\mathbf{y}_1, \ldots, \mathbf{y}_N)} \end{aligned}$$

since $g(T(\mathbf{y}_1, \ldots, \mathbf{y}_N); \theta)$ simplifies in the right hand side's ratio; the latter no longer depends on θ indicating that T is a sufficient statistic. **(Continues...)**

The factorization theorem (4/7)

Theorem 7

Proof.

(Continued.) Consider the "sufficiency" part of the discrete case. Recall the interpretation of a joint p.m.f. as a probability function:

$$\begin{aligned} f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\theta\right) &= \mathbb{P}\left(\bigcup_{i=1}^{N}\boldsymbol{x}_i=\mathbf{x}_i;\theta\right) \\ &= \mathbb{P}\left(\bigcup_{i=1}^{N}\boldsymbol{x}_i=\mathbf{x}_i\middle|T=T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\right) \\ &\quad\times\mathbb{P}\left(T=T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) \\ &= h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)\cdot q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) \end{aligned}$$

The second line follows from the definition of conditional probability while the third just renames the previous probability function, noting that the conditional probability of the sample given T is expressible as some generic function $h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$ that does not depend on θ by the definition of sufficient statistic. **(Continues...)**

The factorization theorem (5/7)

Theorem 7

Proof.

(Continued.) Move to the continuous case. Consider some *bijective* and *differentiable* transformations that *do not depend on* θ:

$$\begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \\ \boldsymbol{g}_2\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \\ \vdots \\ \boldsymbol{g}_N\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right) \end{pmatrix}$$

where at least one element of this list (suppose Y_{11} in $\boldsymbol{y}_1$), is fixed as $Y_{11} = T\left(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\right)$ *by construction*. The *inverse transformation* is as follows.

$$\begin{pmatrix} \boldsymbol{w}_1 \\ \boldsymbol{w}_2 \\ \vdots \\ \boldsymbol{w}_N \end{pmatrix} = \begin{pmatrix} \boldsymbol{g}_1^{-1}\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right) \\ \boldsymbol{g}_2^{-1}\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right) \\ \vdots \\ \boldsymbol{g}_N^{-1}\left(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right) \end{pmatrix}$$

(Continues...)

The factorization theorem (6/7)

Theorem 7

Proof.

(Continued.) In order to show necessity, write the joint p.d.f. of the transformation as:

$$\begin{aligned} f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}\left(\mathbf{y}_1,\ldots,\mathbf{y}_N;\theta\right) &= f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{w}_1,\ldots,\mathbf{w}_N;\theta\right)\cdot\left|\mathbf{J}^*\right| \\ &= g\left(T\left(\mathbf{w}_1,\ldots,\mathbf{w}_N\right);\theta\right)\cdot h\left(\mathbf{w}_1,\ldots,\mathbf{w}_N\right)\cdot\left|\mathbf{J}^*\right| \\ &= g\left(y_{11};\theta\right)\cdot h\left(\mathbf{w}_1,\ldots,\mathbf{w}_N\right)\cdot\left|\mathbf{J}^*\right| \end{aligned}$$

where $|\mathbf{J}^*|$ is shorthand for the absolute value of the Jacobian of the inverse transformation, and the second line follows from hypothesis.

It is obvious that the marginal distribution of Y_{11}, that is the density function $q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)$ of the statistic of interest T, inherits a factorization analogous to the above and since $y_{11}=T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right)$, it can be shown that the ratio between the joint p.d.f. of the sample and the p.d.f. of T does not depend on θ, hence T is sufficient.

(Continues...)

The factorization theorem (7/7)

Theorem 7

Proof.

(Continued.) In order to show the "sufficiency" part of the Theorem (if T is sufficient, then a proper factorization can be expressed) apply the definition of conditional density function to show that:

$$f_{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N}\left(\mathbf{y}_1,\ldots,\mathbf{y}_N;\theta\right) = \\ = q_T\left(y_{11};\theta\right)\cdot f_{\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\}\setminus Y_{11}}\left(\left\{\mathbf{y}_1,\ldots,\mathbf{y}_N\right\}\setminus y_{11}\middle|\, Y_{11}\right)$$

where the notation $\{\cdot\}\setminus Y_{11}$ denotes a list that *excludes* Y_{11}. Dividing both sides of the above by $\left|\mathbf{J}^*\right|$ returns the desired factorization for:

$$h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right) = \frac{f_{\{\boldsymbol{y}_1,\ldots,\boldsymbol{y}_N\}\setminus Y_{11}}\left(\left\{\mathbf{y}_1,\ldots,\mathbf{y}_N\right\}\setminus y_{11}\middle|\, Y_{11}\right)}{\left|\mathbf{J}^*\right|}$$

and for $g\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right) = q_T\left(T\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right);\theta\right)$. □

Uses of the factorization theorem

The factorization theorem is especially useful to show that multiple statistics are **simultaneously** sufficient for a number of associated parameters. This is usually expressed through a *vector* of statistics $\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$.

$$\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \begin{pmatrix} T_1(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ T_2(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \\ \vdots \\ T_K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) \end{pmatrix}$$

These statistics are said to be *simultaneously sufficient* for a *vector of parameters* θ:

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_J \end{pmatrix}$$

where generally it may be that $K \neq J$. The factorization theorem can be extended to allow for $g(\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N); \theta)$ to be the joint p.d.f. of all these statistics and for a multidimensional parameter vector.

Example: sufficiency for the normal distribution

- The earlier result on $\bar{X}$ being sufficient for μ in the normal case can also be obtained via factorization theorem with:

$$g\left(\bar{x};\mu\right)=\exp\left(-\frac{N\left(\bar{x}-\mu\right)^{2}}{2\sigma^{2}}\right)$$

$$h\left(x_{1},\ldots,x_{N}\right)=\left(\frac{1}{2\pi\sigma^{2}}\right)^{\frac{N}{2}}\exp\left(-\sum_{i=1}^{N}\frac{\left(x_{i}-\bar{x}\right)^{2}}{2\sigma^{2}}\right)$$

- . . . but this still ignores σ^2. Consider S^2 and its realization.

$$s^{2}=\frac{1}{N-1}\sum_{i=1}^{N}\left(x_{i}-\bar{x}\right)^{2}$$

- It turns out that $\left(\bar{X},S^{2}\right)$ are jointly sufficient for $\left(\mu,\sigma^{2}\right)$. If

$$g\left(\bar{x},s^{2};\mu,\sigma^{2}\right)=\left(\frac{1}{\sigma^{2}}\right)^{\frac{N}{2}}\exp\left(-\frac{N\left(\bar{x}-\mu\right)^{2}+\left(N-1\right)S^{2}}{2\sigma^{2}}\right)$$

and $h\left(x_{1},\ldots,x_{N}\right)=\left(2\pi\right)^{-N/2}$, the theorem applies nicely.

Example: sufficiency for the multivariate normal

- These results extend to the multivariate case, where $\bar{\boldsymbol{x}}$ with realization $\bar{\mathbf{x}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i$ is sufficient for $\boldsymbol{\mu}$.

$$\frac{f_{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_N}\left(\mathbf{x}_1,\ldots,\mathbf{x}_N;\boldsymbol{\mu},\boldsymbol{\Sigma}\right)}{q_{\bar{\boldsymbol{x}}}\left(\bar{\mathbf{x}};\boldsymbol{\mu},\boldsymbol{\Sigma}/N\right)} = \frac{\exp\left(-\frac{1}{2}\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\right)}{\sqrt{\left[\left(2\pi\right)^{K}\left|\boldsymbol{\Sigma}\right|\right]^{N-1}N}}$$

- Again, this ignores $\boldsymbol{\Sigma}$. Consider $\boldsymbol{S}$ and its realization.

$$\mathbf{S} = \frac{1}{N-1}\sum_{i=1}^{N}\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)\left(\mathbf{x}_i-\bar{\mathbf{x}}\right)^{\mathrm{T}}$$

- Thus, $(\bar{\boldsymbol{x}}, \boldsymbol{S})$ are jointly sufficient for $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Let:

$$g\left(\bar{\mathbf{x}},\mathbf{S};\boldsymbol{\mu},\boldsymbol{\Sigma}\right) =$$
$$= \frac{1}{\left|\boldsymbol{\Sigma}\right|^{\frac{N}{2}}}\exp\left(-\frac{N}{2}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\left(\bar{\mathbf{x}}-\boldsymbol{\mu}\right)-\frac{N-1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{S}\right)\right)$$

and $h\left(\mathbf{x}_1,\ldots,\mathbf{x}_N\right) = \left(2\pi\right)^{-\frac{NK}{2}}$; the theorem applies again.

Example: sufficiency for the uniform distribution

- Let a sample be drawn from $X \sim \mathcal{U}(\alpha, \beta)$, where α and β are unknown parameters.
- The minimum $X_{(1)}$ and maximum $X_{(N)}$, with realizations $x_{(1)} = \min\{x_1, \ldots, x_N\})$ and $x_{(N)} = \max\{x_1, \ldots, x_N\}$, are jointly sufficient for α and β.
- The joint p.d.f. of the sample here is:

$$f_{X_1,\ldots,X_N}(x_1, \ldots, x_N; \alpha, \beta) = \\ = \left(\frac{1}{\beta - \alpha}\right)^N \cdot \mathbb{1}[\alpha \leq x_1, \ldots, x_N \leq \beta]$$

- ... hence, the factorization theorem here applies by setting

$$g\left(x_{(1)}, x_{(N)}; \alpha, \beta\right) = \left(\frac{1}{\beta - \alpha}\right)^N \cdot \mathbb{1}\left[\alpha \leq x_{(1)}\right] \cdot \mathbb{1}\left[x_{(N)} \leq \beta\right]$$

and $h(x_1, \ldots, x_N) = 1$.

The exponential (macro-)family

The factorization theorem is extremely easy to apply to a wide array of distributions.

Definition 20

Exponential (Macro-)family. A family of probability distributions expressed by a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_J)$ is said to belong to the *exponential* (macro)-family if the associated p.m.f.s or p.d.f.s can be written, for $J \leq L$, as follows.

$$f_X(x; \boldsymbol{\theta}) = h(x)\, c(\boldsymbol{\theta}) \exp\left(\sum_{\ell=1}^{L} w_\ell(\boldsymbol{\theta})\, t_\ell(x)\right)$$

Here $h(x)$ and $t_\ell(x)$ are functions of the realizations x, $c(\boldsymbol{\theta}) \geq 0$ and $w_\ell(\boldsymbol{\theta})$ are functions of the parameters $\boldsymbol{\theta}$; with $\ell = 1, \ldots, L$.

- The families: Bernoulli, geometric, Poisson, normal, lognormal, Beta, Gamma (including its special cases) are all sub-families of the exponential macro-family.

Sufficiency and the exponential family

Theorem 8

Sufficient statistics and the exponential family. *If a random sample is obtained from any random variable X whose distribution belongs to the exponential family, the L statistics in the vector:*

$$\boldsymbol{t}\left(X_1,\ldots,X_N\right)=\begin{pmatrix}\sum_{i=1}^{N} t_1\left(X_i\right)\\ \sum_{i=1}^{N} t_2\left(X_i\right)\\ \vdots\\ \sum_{i=1}^{N} t_L\left(X_i\right)\end{pmatrix}$$

are simultaneously sufficient for θ, where the functions $t_\ell(x)$ are as in the previous definition of the exponential family for $\ell=1,\ldots,L$.

Proof.

The joint density of the sample can be expressed as:

$$f_{X_1,\ldots,X_N}\left(x_1,\ldots,x_N;\theta\right)=\left(\prod_{i=1}^{N} h\left(x_i\right)\right)\left[c\left(\theta\right)\right]^N\exp\left(\sum_{\ell=1}^{L} w_\ell\left(\theta\right)\sum_{i=1}^{N} t_\ell\left(x_i\right)\right)$$

and applying the factorization theorem is straightforward. □

Sufficiency and the exponential family: examples

- The Bernoulli p.m.f. can be written, for $x \in \{0, 1\}$, as:

$$f_X(x, p) = (1 - p) \exp\left(\log\left(\frac{p}{1-p}\right) x\right)$$

  implying that $T = \sum_{i=1}^{N} X_i$ is sufficient for p.

- The normal p.d.f. can be written as:

$$f_X\left(x; \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \exp\left(\frac{\mu}{\sigma^2} x - \frac{1}{2\sigma^2} x^2\right)$$

  so $T_1 = \sum_{i=1}^{N} X_i$ and $T_2 = \sum_{i=1}^{N} X_i^2$ "suffice" for μ and σ^2.

- The Gamma p.d.f. can be written, for $x > 0$, as:

$$f_X(x; \alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \exp\left[(\alpha - 1) \log(x) - \beta x\right]$$

  so $T_1 = \sum_{i=1}^{N} \log(X_i)$ and $T_2 = \sum_{i=1}^{N} X_i$ suffice for α and β.

Transformations of sufficient statistics

- If $T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ is a sufficient statistic for some parameter θ, a **transformation** $T'(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = g(T(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$ is also sufficient for θ if $g(\cdot)$ does not depend on θ.

- This conclusion also applies to the **multidimensional** case where $\boldsymbol{t}'(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \boldsymbol{g}(\boldsymbol{t}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N))$.

- Example (normal): if $T_1 = \sum_{i=1}^{N} X_i$ & $T_2 = \sum_{i=1}^{N} X_i^2$:

$$\bar{X} = \frac{1}{N} T_1 \quad \text{and} \quad S^2 = \frac{1}{N-1}\left(T_2 - \frac{T_1^2}{N}\right)$$

  are also sufficient for μ and σ^2 of the normal distribution.

- Example (Gamma): if $T_1 = \sum_{i=1}^{N} \log(X_i)$ & $T_2 = \sum_{i=1}^{N} X_i$:

$$T_1' = \exp(T_1) = \prod_{i=1}^{N} X_i \quad \text{and} \quad T_2' = T_2 = \sum_{i=1}^{N} X_i$$

  are also sufficient for α and β of the Gamma distribution.