S&DS 242/542: Theory of Statistics

Lecture 4: Large-sample approximations, computer simulation

Sampling distributions of statistics

For data $X_1, \ldots, X_n$, a **statistic** $T(X_1, \ldots, X_n)$ is any function of the data. For example:

$$\begin{aligned}
\text{(sample mean) } \bar{X} &= \frac{X_1 + \ldots + X_n}{n} \\
\text{(sample variance) } S^2 &= \frac{(X_1 - \bar{X})^2 + \ldots + (X_n - \bar{X})^2}{n-1} \\
\text{(sample range) } R &= \max(X_1, \ldots, X_n) - \min(X_1, \ldots, X_n)
\end{aligned}$$

If the data $X_1, \ldots, X_n$ are random, then this randomness induces a **sampling distribution** for the statistic.

If we understand the randomness of $X_1, \ldots, X_n$, how can we understand the sampling distribution of $T(X_1, \ldots, X_n)$?

Example: Sample mean of normally-distributed data

Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \mathcal{N}(\mu, \sigma^2)$. What is the distribution of the sample mean

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}?$$

We have

$$\mathbb{E}[\bar{X}] = \frac{\mathbb{E}X_1 + \ldots + \mathbb{E}X_n}{n} = \mu$$

$$\operatorname{Var}[\bar{X}] = \frac{\operatorname{Var} X_1 + \ldots + \operatorname{Var} X_n}{n^2} = \frac{\sigma^2}{n}.$$

From last lecture, any linear combination of independent normal variables has a normal distribution. So

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

The need for approximation

For many (seemingly simple) statistics, its sampling distribution is difficult to describe exactly. For example:

1. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Uniform}(-1, 1)$. What is the distribution of the sample mean $\bar{X}$?
2. Suppose $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\frac{1}{2})$, the outcomes of n tosses of a fair coin. What is the distribution of

   $$T = \left(\bar{X} - \frac{1}{2}\right)^2?$$

   If we compute T from n observed coin tosses and this is too large compared with typical values from this distribution, then we have evidence that the coin may not be fair. (This is the idea of *hypothesis testing*, which we'll discuss next class.)

For distributions that are difficult to study exactly, we can try to study them via *computer simulation* or *large-sample approximation*.

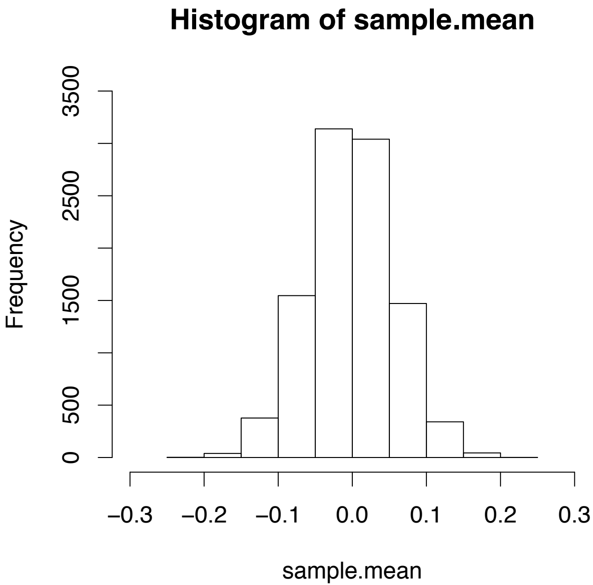
Sample mean of IID uniform

If we know the distribution of the data, then we can **simulate** the distribution of any statistic computed from this data.

Here is a snippet of R code that simulates the distribution of the sample mean $\bar{X}$ of 100 uniform random variables $X_1, \ldots, X_{100} \overset{IID}{\sim} \text{Uniform}(-1, 1)$:

```
sample.mean = numeric(10000)
for (i in 1:10000) {
    X = runif(100, min=-1, max=1)
    sample.mean[i] = mean(X)
}
hist(sample.mean)
```

Sample mean of IID uniform



Is the coin fair?

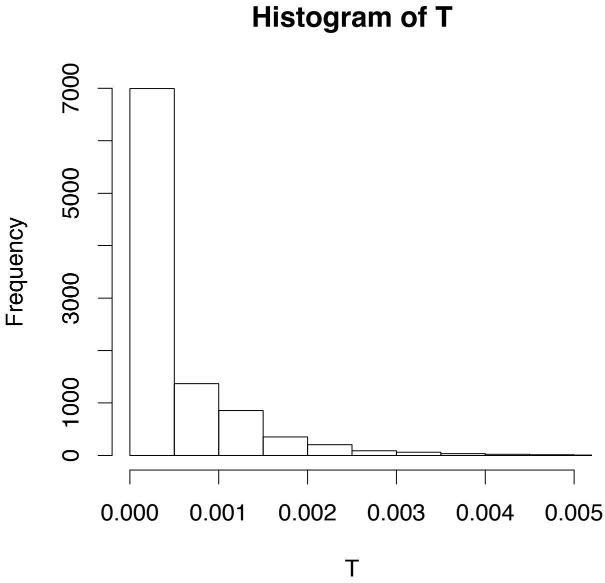
Here is a snippet of `R` code that plots a histogram of the statistic

$$T = \left(\bar{X} - \frac{1}{2}\right)^2$$

where $\bar{X}$ is the fraction of heads in 500 tosses of a fair coin.

```
T = numeric(10000)
for (i in 1:10000) {
    S = rbinom(1, size=500, prob=1/2)
    T[i] = (S/500 - 1/2)^2
}
hist(T)
```

Is the coin fair?



Approximating probabilities and expectations

Suppose we simulate b values $T_1, \ldots, T_b$ according to the distribution of T. If we're interested in $\mathbb{P}[T \in A]$ for some set $A \subseteq \mathbb{R}$, we may approximate

$$\mathbb{P}[T \in A] \approx \frac{\#\text{ simulations } i \text{ where } T_i \in A}{b}$$

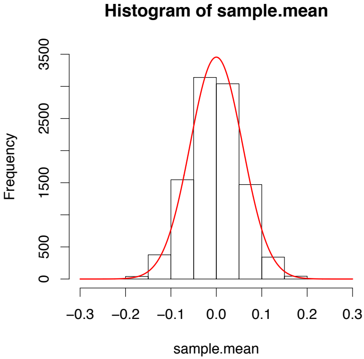
If we're instead interested in $\mathbb{E}[f(T)]$ for some function $f : \mathbb{R} \to \mathbb{R}$, we may approximate

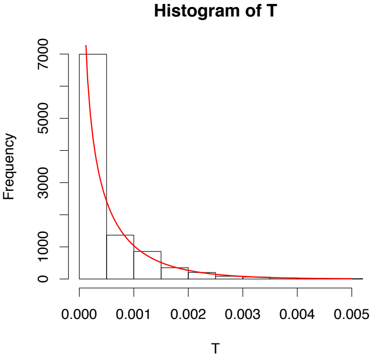
$$\mathbb{E}[f(T)] \approx \frac{1}{b} \sum_{i=1}^{b} f(T_i).$$

These approximations become more accurate as the number of simulations b increases. These are the simplest examples of **Monte Carlo approximations**.

Large-sample approximations

Oftentimes, a good approximation to the distribution of $T = T(X_1, \ldots, X_n)$ emerges when the sample size n is large. We call such results **asymptotic** or **large-sample approximations**.





Large-sample approximations

Many statistical procedures (for example, tests of independence for categorical data or confidence intervals for logistic regression) are based on asymptotic approximations.

In the computer age, some of this need for asymptotic approximations has been supplanted by the ease of simulation.

Here are two reasons why we still use asymptotic approximations:

1. It's (much) faster to get an answer.
2. It's useful to have theoretical understanding.
   - What if the X_i's are not actually Uniform$(-1, 1)$? What if I don't really know the true distribution of the X_i's?
   - What if $n = 1000$ instead of 100? $n = 1{,}000{,}000$ instead of 100? What n do I need so that my estimation error is $< 1\%$?

(Weak) Law of Large Numbers

The large-sample approximations in this course will be based on two fundamental results from probability theory: The **(Weak) Law of Large Numbers** and the **Central Limit Theorem**.

Theorem (Weak Law of Large Numbers)

Suppose $X_1, \ldots, X_n$ *are IID, with* $\mathbb{E}[X_1] = \mu$ *and* $\mathrm{Var}[X_1] < \infty$. *Let*

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

Then $\bar{X} \to \mu$ *in probability, as* $n \to \infty$.

This means: For any fixed interval $(\mu - \varepsilon, \mu + \varepsilon)$ around μ, where $\varepsilon > 0$ can be arbitrarily small (not depending on n), the probability that $\bar{X}$ belongs to $(\mu - \varepsilon, \mu + \varepsilon)$ approaches 1 as $n \to \infty$.

(Weak) Law of Large Numbers

Example

Let $X_1, \ldots, X_n \overset{IID}{\sim}$ Bernoulli($\frac{1}{2}$) represent n tosses of a fair coin, where $X_i = 1$ for heads and $X_i = 0$ for tails. Then

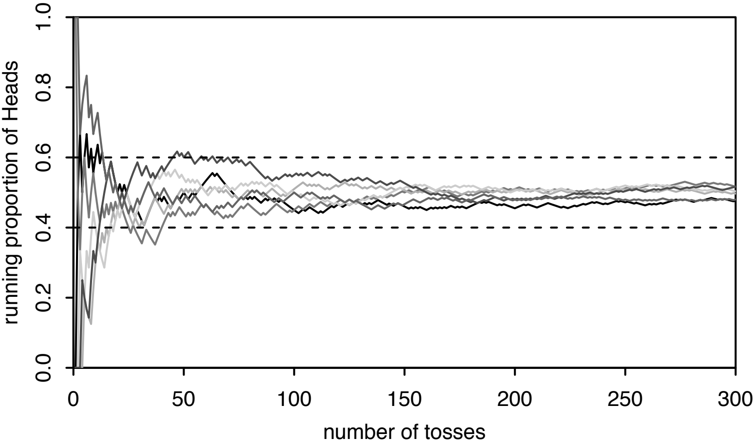
$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

is the fraction of heads among these n tosses.

Consider the interval $(0.4, 0.6)$ around 0.5. For each value of n, there is some probability that $\bar{X}$ falls outside $(0.4, 0.6)$. The LLN guarantees that as $n \to \infty$, this probability goes to 0, and the probability that $\bar{X}$ belongs to $(0.4, 0.6)$ goes to 1.

The same guarantee holds for any fixed interval around 0.5: $(0.45, 0.55)$, $(0.49, 0.51)$, $\ldots$

(Weak) Law of Large Numbers



Central Limit Theorem

How close to μ should we expect $\bar{X}$ to actually be? What is the distribution of $\bar{X}$ around μ?

Theorem (Central Limit Theorem)

Suppose $X_1, \ldots, X_n$ are IID, with $\mathbb{E}[X_1] = \mu$ and $\mathrm{Var}[X_1] = \sigma^2$. Let $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$. Then

$$\sqrt{n}\left(\frac{\bar{X} - \mu}{\sigma}\right) \to \mathcal{N}(0, 1)$$

in distribution, as $n \to \infty$. Equivalently,

$$\sqrt{n}\,(\bar{X} - \mu) \to \mathcal{N}(0, \sigma^2)$$

This means: For any fixed interval (a, b), the probability that $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma})$ belongs to (a, b) approaches the probability that a standard normal variable $Z \sim \mathcal{N}(0, 1)$ belongs to (a, b), as $n \to \infty$.

Central Limit Theorem

More informally, when the sample size n is large:

- The distribution of $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma})$ is approximately $\mathcal{N}(0,1)$.

  If we simulate the value of $\bar{X}$ many times, the histogram of the values $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma})$ will be close in shape to the standard normal bell curve.
- The distribution of $\bar{X}$ is approximately $\mathcal{N}(\mu, \frac{\sigma^2}{n})$.
  This holds even when $X_1, \ldots, X_n$ are not normally distributed.

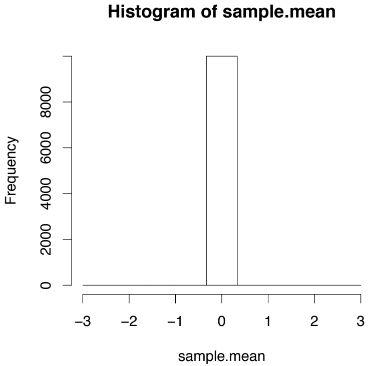
[Note: It's formally not correct to say that

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

as $n \to \infty$, because this limit $\mathcal{N}(\mu, \frac{\sigma^2}{n})$ depends on n. This is why we write instead $\sqrt{n}(\frac{\bar{X}-\mu}{\sigma}) \to \mathcal{N}(0,1)$ or $\sqrt{n}(\bar{X}-\mu) \to \mathcal{N}(0,\sigma^2)$.]

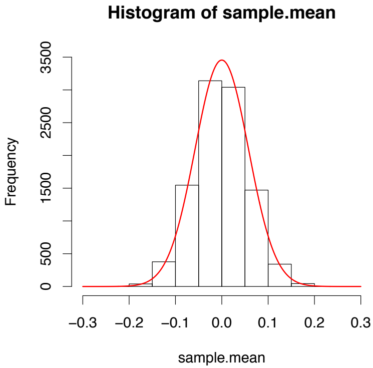
The difference is in the scaling

The LLN describes the behavior of $\bar{X}$ on the "constant scale", saying that for any interval around μ of *constant size* (not depending on *n*), $\bar{X}$ belongs to this interval with high probability. Here is a histogram of simulations of $\bar{X}$:



The difference is in scaling

The CLT describes the behavior of $\bar{X}$ on the "$\frac{1}{\sqrt{n}}$ scale", saying that for an interval around μ whose length is on the order of $\frac{1}{\sqrt{n}}$, the probability that $\bar{X}$ belongs to this interval is approximately the area under a normal bell curve. Here is the same histogram, zoomed in to a smaller scale:



Sample mean of IID uniform

Example

Let $X_1, \ldots, X_n \overset{IID}{\sim} \text{Uniform}(-1, 1)$. We have $\mathbb{E}[X_1] = 0$, and

$$\text{Var}[X_1] = \mathbb{E}[X_1^2] = \int_{-1}^{1} x^2 \cdot \frac{1}{2}\, dx = \frac{1}{3}.$$

By the LLN,

$$\bar{X} \to \mathbb{E}[X_1] = 0$$

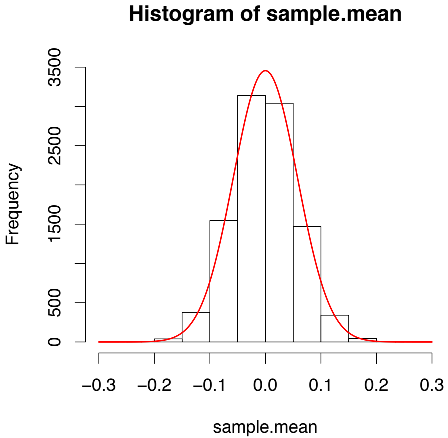
in probability as $n \to \infty$.

By the CLT,

$$\sqrt{3n} \cdot \bar{X} \to \mathcal{N}(0, 1) \quad \text{or} \quad \sqrt{n} \cdot \bar{X} \to \mathcal{N}(0, \tfrac{1}{3})$$

in distribution. More informally, the distribution of $\bar{X}$ is approximately $\mathcal{N}(0, \frac{1}{3n})$ for large n.

Sample mean of IID uniform



The red curve corresponds to the PDF of $\mathcal{N}(0, \frac{1}{3n})$, for $n = 100$.

Sample mean of IID uniform

Th CLT tells us that the distribution of $\bar{X}$ is approximately $\mathcal{N}\left(0, \frac{1}{3n}\right)$. How good is this approximation?

Here's a comparison of CDF values, for sample size $n = 10$:[1]

Normal Approximation	Exact
0.01	0.009
0.25	0.253
0.50	0.500
0.75	0.747
0.99	0.991

It's already very close! In general, the accuracy depends on

- Sample size n
- Skewness of the distribution of the X_i's
- Heaviness of tails of the distribution of the X_i's

[1] Using `www.math.uah.edu/stat/apps/SpecialCalculator.html`

Central Limit Theorem

Example

Let's continue the example $X_1, \ldots, X_n \overset{IID}{\sim} \text{Bernoulli}(\frac{1}{2})$. We have $\mathbb{E}[X_1] = \frac{1}{2}$ and $\text{Var}[X_1] = \frac{1}{4}$. The CLT tells us that

$$\sqrt{4n} \cdot (\bar{X} - \tfrac{1}{2}) \to \mathcal{N}(0, 1) \quad \text{or} \quad \sqrt{n} \cdot (\bar{X} - \tfrac{1}{2}) \to \mathcal{N}(0, \tfrac{1}{4})$$

in distribution. More informally, the distribution of $\bar{X}$ is approximately $\mathcal{N}(\frac{1}{2}, \frac{1}{4n})$ for large n.

Is the coin fair?

Consider our previous example, the statistic

$$T = \left(\bar{X} - \frac{1}{2}\right)^2.$$

Since the distribution of $\bar{X} - \frac{1}{2}$ is approximately normal, is the distribution of T approximately the square of a normal?

Yes! This is guaranteed by the **Continuous Mapping Theorem**.

Theorem (Continuous Mapping)

Let $g(x)$ be a continuous function of x. As $n \to \infty$,

(a) *If $T_n \to Z$ in distribution, then $g(T_n) \to g(Z)$ in distribution.*

(b) *If $T_n \to \mu$ in probability, then $g(T_n) \to g(\mu)$ in probability.*

Is the coin fair?

By the CLT,

$$\sqrt{4n} \cdot (\bar{X} - \tfrac{1}{2}) \to \mathcal{N}(0, 1).$$

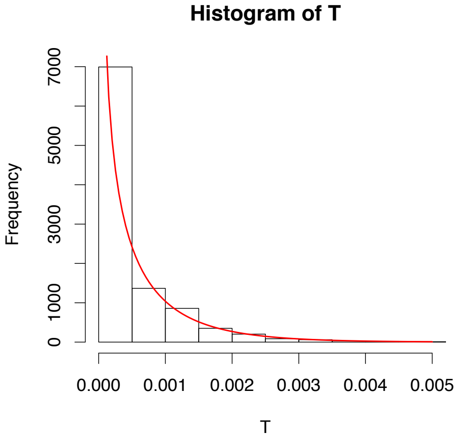
Then by the Continuous Mapping Theorem,

$$4n \cdot (\bar{X} - \tfrac{1}{2})^2 \to \chi_1^2.$$

Recall that χ_1^2 is the distribution of Z^2 when $Z \sim \mathcal{N}(0, 1)$.

More informally, for large n, we'll say that the distribution of $(\bar{X} - \frac{1}{2})^2$ is approximately $\frac{1}{4n} \cdot \chi_1^2$.

Is the coin fair?



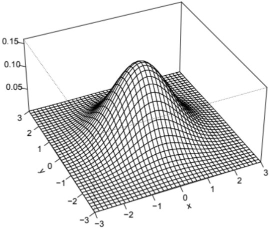
The red curve corresponds to the PDF of $\frac{1}{4n} \cdot \chi_1^2$, for $n = 500$.

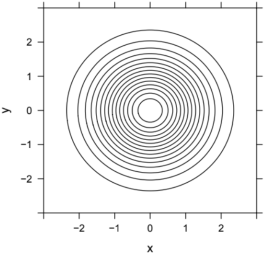
The standard multivariate normal

Recall that the **standard multivariate normal** distribution in $\mathbb{R}^k$ is the joint distribution of $X_1, \ldots, X_k \overset{IID}{\sim} \mathcal{N}(0,1)$, with joint PDF

$$f(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{k/2}} e^{-\frac{1}{2}(x_1^2 + \ldots + x_k^2)}$$

This PDF is symmetric under rotations/reflections about the origin.

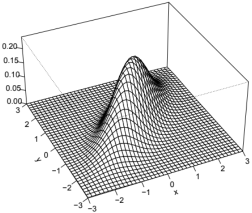


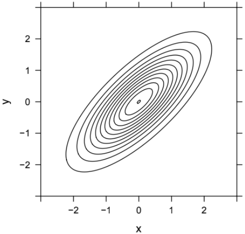


The general multivariate normal

For general mean vector $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$, the **multivariate normal** distribution $(X_1, \ldots, X_k) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is a continuous distribution with joint PDF

$$f(\mathbf{x}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Sigma})}} \, e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \text{ where } \mathbf{x} = (x_1, \ldots, x_k)$$





- Each coordinate X_i has mean $\mathbb{E}[X_i] = \mu_i$
- Each coordinate X_i has variance $\mathrm{Var}[X_i] = \Sigma_{ii}$
- Each coordinate pair X_i, X_j has covariance $\mathrm{Cov}[X_i, X_j] = \Sigma_{ij}$

Representation by standard multivariate normal

An alternative way to define the distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is via the representation $X = \mu + \sigma Z$ where $Z \sim \mathcal{N}(0, 1)$. Analogously:

Theorem

If $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ *is any multivariate normal vector in* $\mathbb{R}^k$, *then*

$$\mathbf{X} = \boldsymbol{\mu} + \mathbf{V}\mathbf{Z}$$

for a standard multivariate normal vector $\mathbf{Z} \in \mathbb{R}^k$ *and some matrix* $\mathbf{V} \in \mathbb{R}^{k \times k}$. *This matrix* $\mathbf{V}$ *must satisfy* $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{V}^\top$.

Conversely, if $\mathbf{X} = \boldsymbol{\mu} + \mathbf{V}\mathbf{Z}$ *where* $\boldsymbol{\mu}, \mathbf{V}$ *are any fixed vector/matrix and* $\mathbf{Z}$ *is a standard multivariate normal vector, then*

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \textit{where} \quad \boldsymbol{\Sigma} = \mathbf{V}\mathbf{V}^\top.$$

This gives an alternative definition of $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Individual coordinates and their linear combinations

Suppose $\mathbf{X} = (X_1, \ldots, X_k)$ has a multivariate normal distribution. Let $\mathbf{a} = (a_1, \ldots, a_k) \in \mathbb{R}^k$ be any fixed vector. Then

$$\mathbf{a}^\top \mathbf{X} = a_1 X_1 + \ldots + a_k X_k$$

has a normal distribution.

Proof: Represent $X = \mu + VZ$, where $\underbrace{Z \sim \mathcal{N}(0, I)}_{\text{standard MVN}}$

$$\text{So } a^T X = \underbrace{a^T \mu}_{=c} + \underbrace{a^T V}_{=b^T} Z = c + b^T Z$$

This is normal from last class.

In particular, coordinates of $\mathbf{X}$ have distributions $X_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$.

Linear transformations

Suppose **X** has a multivariate normal distribution. Let **a** and **B** be any fixed vector/matrix. Then

$$\mathbf{Y} = \mathbf{a} + \mathbf{B}\mathbf{X}$$

also has a multivariate normal distribution.

Proof: Represent $X = \mu + VZ$, where $Z \sim N(0, I)$

So $Y = a + BX$

$= a + B(\mu + VZ) = \underbrace{a + B\mu}_{= \mu'} + \underbrace{BV}_{= V'} Z$

$= \mu' + V'Z$. So this is multivariate normal

Multivariate LLN and CLT

The Law of Large Numbers and Central Limit Theorem extend to a multivariate setting:

Let $\mathbf{X} \in \mathbb{R}^k$ be a random vector with mean $\boldsymbol{\mu} \in \mathbb{R}^k$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{k \times k}$. This means

$$\mathbb{E}[X_i] = \mu_i, \quad \mathrm{Var}[X_i] = \Sigma_{ii}, \quad \mathrm{Cov}[X_i, X_j] = \Sigma_{ij} \text{ for all } i \neq j.$$

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n \in \mathbb{R}^k$ be IID random vectors with the same distribution as $\mathbf{X}$. Consider $\bar{\mathbf{X}} = \frac{1}{n}(\mathbf{X}_1 + \ldots + \mathbf{X}_n) \in \mathbb{R}^k$.

Theorem (LLN)

As $n \to \infty$, $\bar{\mathbf{X}}$ converges in probability to $\boldsymbol{\mu}$.

Theorem (CLT)

As $n \to \infty$, $\sqrt{n}\,(\bar{\mathbf{X}} - \boldsymbol{\mu})$ converges in distribution to the multivariate normal $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$.

Example of multivariate CLT

Example

Consider n people sampled independently (with replacement) from a population. Let X_i be the height and Y_i the weight of person i. Note that X_i may be correlated with Y_i, but the pairs (X_i, Y_i) are IID across different people $i = 1, \ldots, n$.

Let $\bar{X} = \frac{1}{n}(X_1 + \ldots + X_n)$ and $\bar{Y} = \frac{1}{n}(Y_1 + \ldots + Y_n)$ be their average height and average weight.

If $\mathbb{E}[X_1] = \mu_X$ and $\mathbb{E}[Y_1] = \mu_Y$, then the LLN tells us that

$$\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} \to \boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \text{in probability, as } n \to \infty.$$

This means: For any fixed ball around (μ_X, μ_Y) in the plane, the probability $(\bar{X}, \bar{Y})$ belongs to this ball approaches 1, as $n \to \infty$.

Multivariate generalizations

Example (Cont'd)

Suppose $\mathrm{Var}[X_1] = \sigma_X^2$, $\mathrm{Var}[Y_1] = \sigma_Y^2$, and $\mathrm{Cov}[X_1, Y_1] = \rho\sigma_X\sigma_Y$. Here σ_X is the standard deviation of height, σ_Y is the standard deviation of weight, and ρ is the correlation between height and weight. Consider the mean vector and 2×2 covariance matrix

$$\boldsymbol{\mu} = \begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_X^2 & \rho\sigma_X\sigma_Y \\ \rho\sigma_X\sigma_Y & \sigma_Y^2 \end{pmatrix}$$

The CLT tells us that

$$\sqrt{n}\left(\begin{pmatrix} \bar{X} \\ \bar{Y} \end{pmatrix} - \boldsymbol{\mu}\right) \to \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{in distribution, as } n \to \infty.$$

Informally: The distribution of $(\bar{X}, \bar{Y})$ is approximately $\mathcal{N}(\boldsymbol{\mu}, \frac{1}{n}\boldsymbol{\Sigma})$. Since height and weight are correlated, the average height $\bar{X}$ and average weight $\bar{Y}$ remain correlated in this normal approximation.