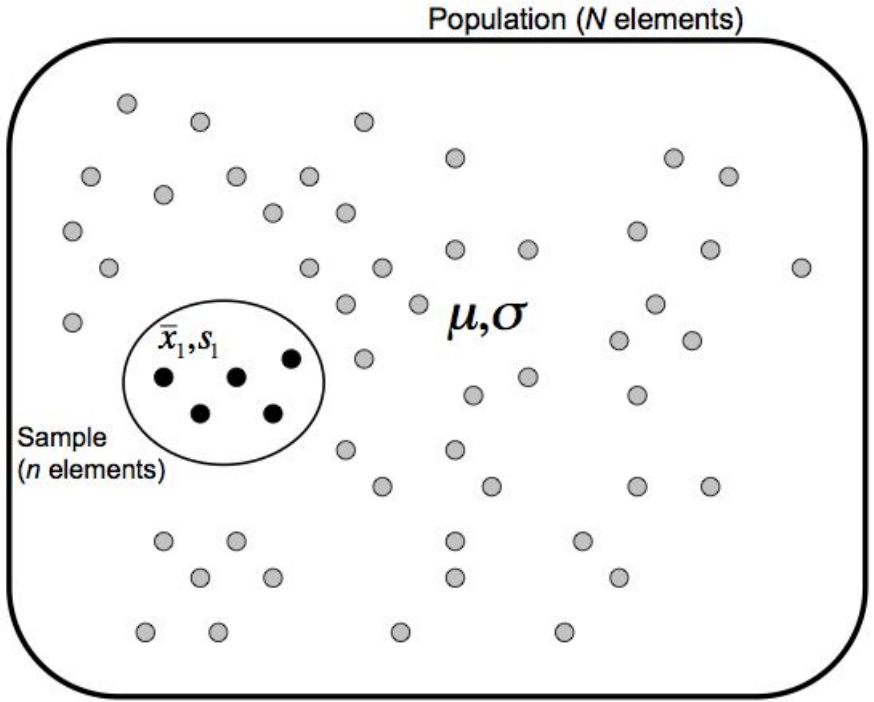
Chapitre 4. Des populations aux échantillons et retour : échantillonnage et estimation

Andreas Zanzoni (orcid.org/0000-0002-4818-6161)
Jacques van Helden (orcid.org/0000-0002-8799-8584)
Lou BERGOGNE (orcid.org/0009-0004-9409-0154)

Estimating a parameter of the population from a sample

- Let us consider a **population** of *N* elements, with *N* assumed to be large.
- We draw a **sample** of *n* elements from the population.
- For each element of this sample, we measure a certain **quantitative property**, and we obtain the **sample values**.

$$\{x_1, x_2, \ldots, x_n\}$$



- Based on the sample, we want to **estimate certain parameters of the population**, such as population mean or population standard deviation.

Population mean (unknown)	μ
Population sd (unknown)	σ
Sample	$\{x_1, x_2, \ldots, x_n\}$
Sample mean	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
Sample sd	$s = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Estimating a parameter of the population from a sample

Situation

- We have a **random sample** drawn from a population.
- We can easily compute some ***sample statistics*** such as sample mean and sample standard deviation.
- From these, we would like to ***estimate*** the corresponding ***population parameters***.

Problem

- Different samples drawn from the same population will generally produce different values of the sample mean and sample standard deviation. These sample statistics are therefore **random variables**: their values vary from sample to sample.

$\overline{x}, S$

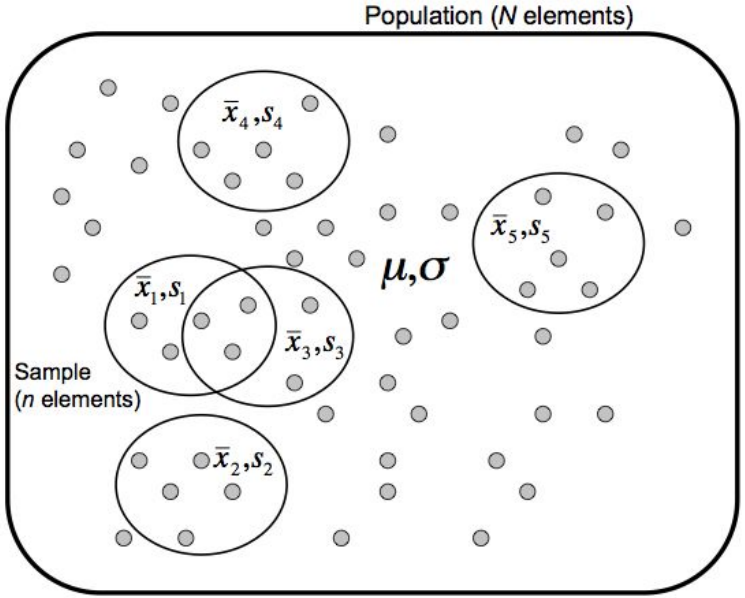
- In contrast, the **population parameters** (such as the population mean and population standard deviation) are **fixed constants**, even though we do not know their true values.

μ, σ

Question

To what extent can we rely on the mean and standard deviation of the sample to estimate the mean and standard deviation of the population ?
In other words:

- *How close are the sample statistics likely to be to the true population parameters?*
- *How large is the sampling variability?*
- *How does the sample size (n) affect the precision and reliability of these estimates?*



Expectation

Discrete variables

$$E(Y) = \sum_{x \in D} P(x)y(x)$$

Continuous variables

$$E(Y) = \int_{-\infty}^{+\infty} f(x)y(x)dx$$

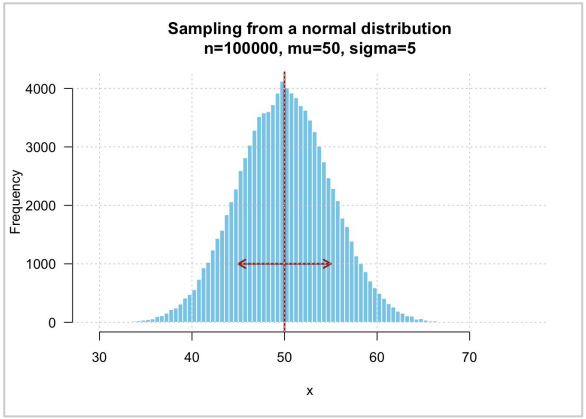
- Where
  - *P(x)* is the probability to observe the value *x* (discrete random variables)
  - *f(x)* is the density function (continuous random variables)
  - *f(x)dx* is an element of probability
  - *y(x)* can be any function of the random distribution x,
    The mean, the variance, the median, ...
- *E(Y)* is called the ***expectation*** for the random variable Y defined by the function *y(x)*.

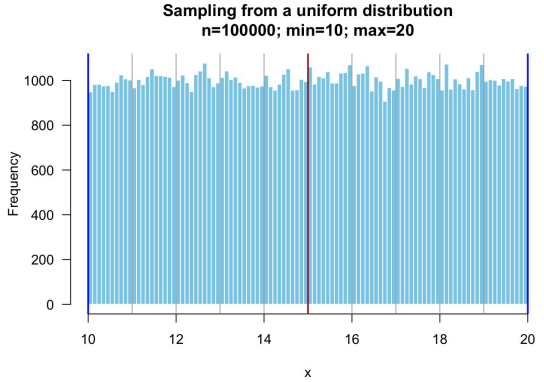
Tirages aléatoires dans une distribution théorique

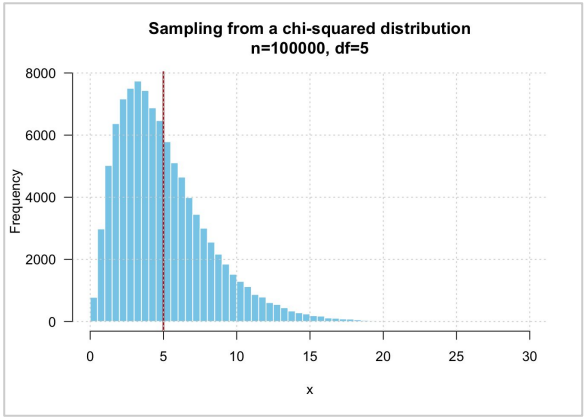
Simulation numérique : nous tirons *n* nombres aléatoires dans une distribution théorique donnée. Chaque distribution est définie par ses paramètres.

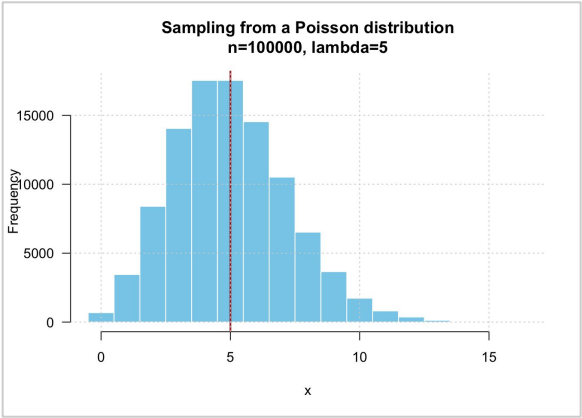
Quelques exemples (parmi bien d'autres)

- Normale: moyenne (μ=50) et écart-type (σ=5)
- Uniforme : bornes min=10 et max=20
- Chi carrée : degrés de liberté (df=5)
- Poisson: espérance (λ=5)









Distribution d'échantillonnage de la moyenne

Chaque fois que nous tirons un échantillon aléatoire d'une population, nous obtenons des valeurs différentes, dont la moyenne fluctue donc d'un tirage à l'autre.

Simulation numérique

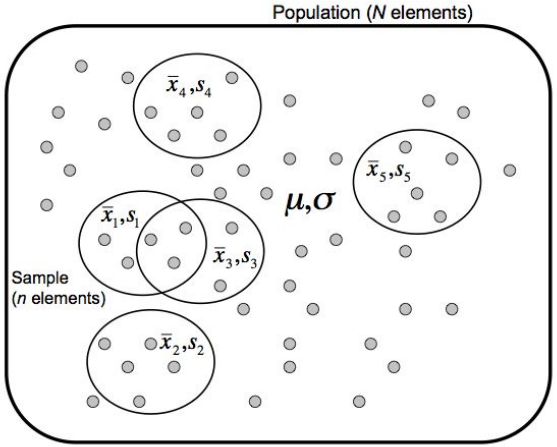
- Nous répétons un grand nombre de fois (r = 100.000) un tirage aléatoire de 5 nombres dans une distribution théorique **uniforme**, bornée de +10 à +20.
- Après chaque tirage, nous calculons la moyenne de cet échantillon-là.
- Nous dessinons ensuite la distribution de toutes les 100.000 moyennes d'échantillons. Cette distribution est appelée la **distribution d'échantillonnage de la moyenne** (***sampling mean distribution***).

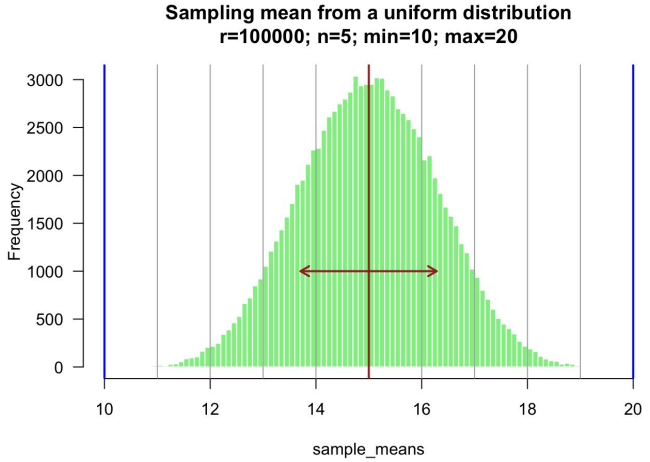
Observations

- La distribution d'échantillonnage de la moyenne suit une courbe en cloche (alors que notre distribution d'origine était uniforme).
- Cette courbe est centrée sur 15, qui est la moyenne théorique d'une distribution uniforme allant de +10 à +20.

Question

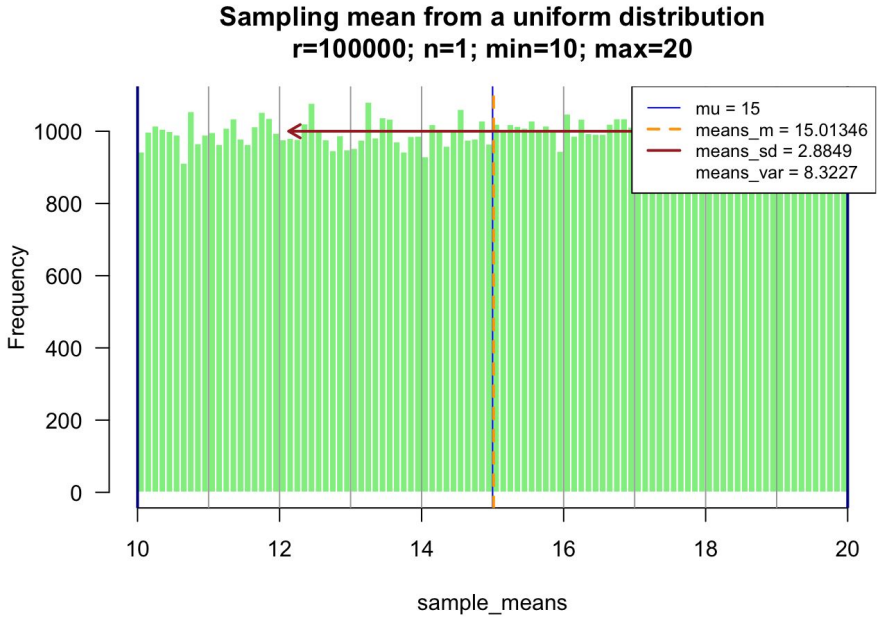
- Que se passe-t-il si l'on refait cette expérience (100.000 répétitions de calcul d'une moyenne d'échantillon) en modifiant la taille de l'échantillon (par exemple n = 1, 2, 4, 8, 16, 32, 64, …) ?





Distribution de la moyenne d'échantillons de n=1 nombre (cas trivial)

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons ne comportant chacun qu'un seul élément (***n=1***) tiré dans une distribution uniforme s' étendant de +10 à +20.
- En toute logique, la distribution de la moyenne d' échantillons suit une distribution uniforme, avec les mêmes bornes de +10 à +20.
- Notons que la variance de ces moyennes vaut 8.32

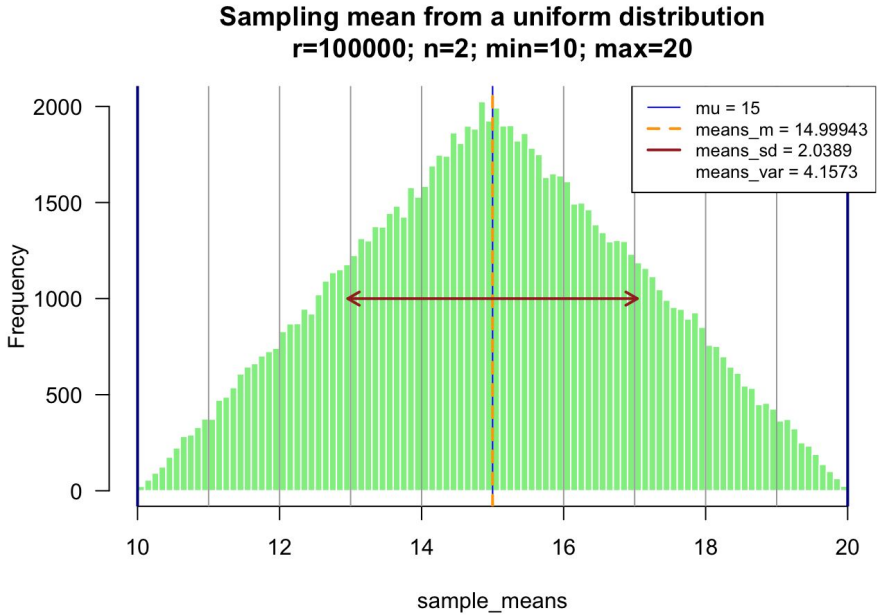


Distribution de la moyenne d'échantillons de n=2 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=2*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme de triangle.

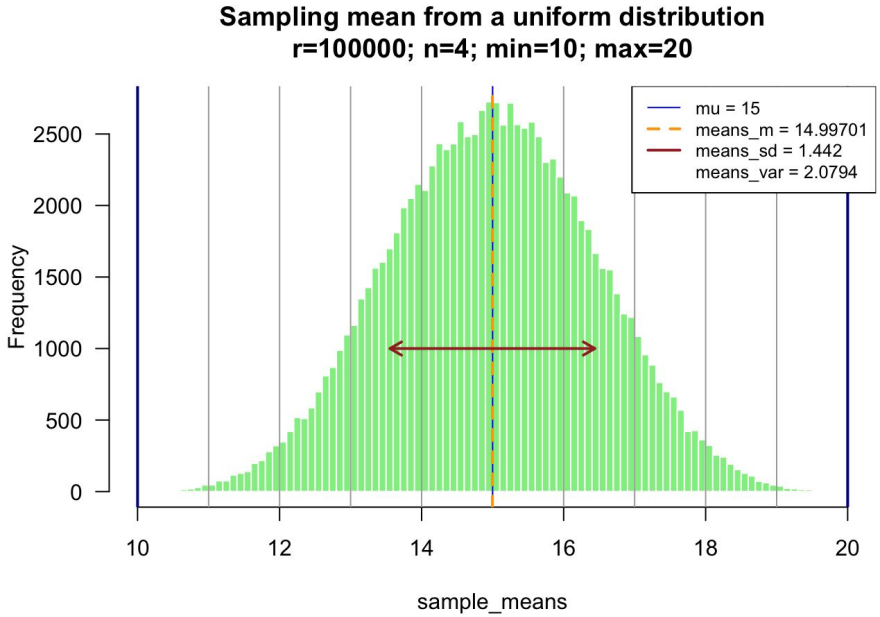
Comment comprendre cela ?

- Les valeurs centrales peuvent s'obtenir de plusieurs façons : en tirant deux nombres autour de la moyenne (15, et 15), ou bien deux nombres dont les différences se compensent (14 et 16, 13 et 17, 12 et 18, …).
- Par contre, il y a moins de chance de tirer par hasard deux nombres très proches du minimum (10 et 10) ou du maximum (20 et 20).
- Note: la variance de ces moyennes de 2 nombres vaut 4.15



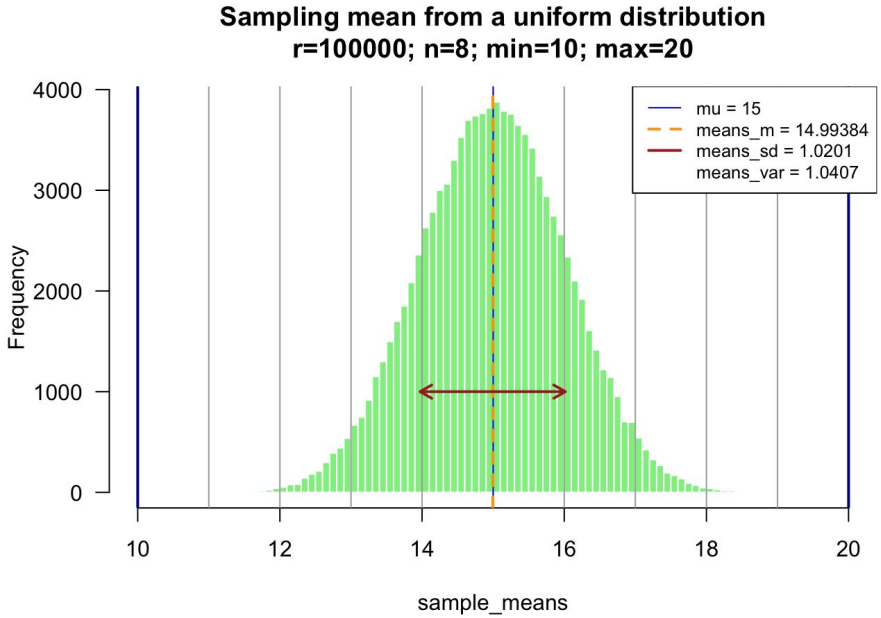
Distribution de la moyenne d'échantillons de n=4 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=4*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme en cloche, centrée sur 15 (la moyenne de la distribution uniforme).
- Note: la variance de ces moyennes de 4 nombres vaut 2.08.



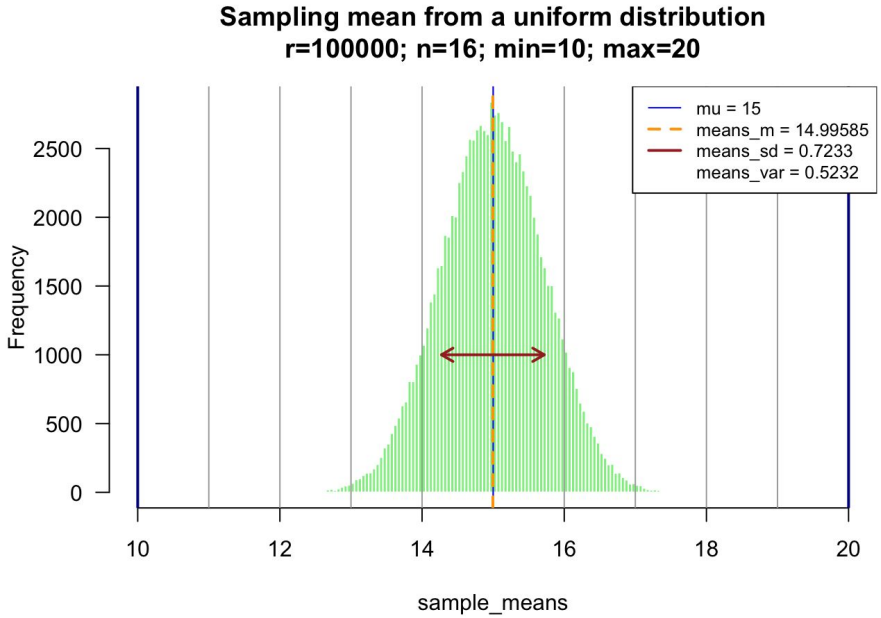
Distribution de la moyenne d'échantillons de n=8 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=8*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme en cloche, centrée sur 15 (la moyenne de la distribution uniforme).
- Note: la variance de ces moyennes de 5 nombres vaut 1.04.



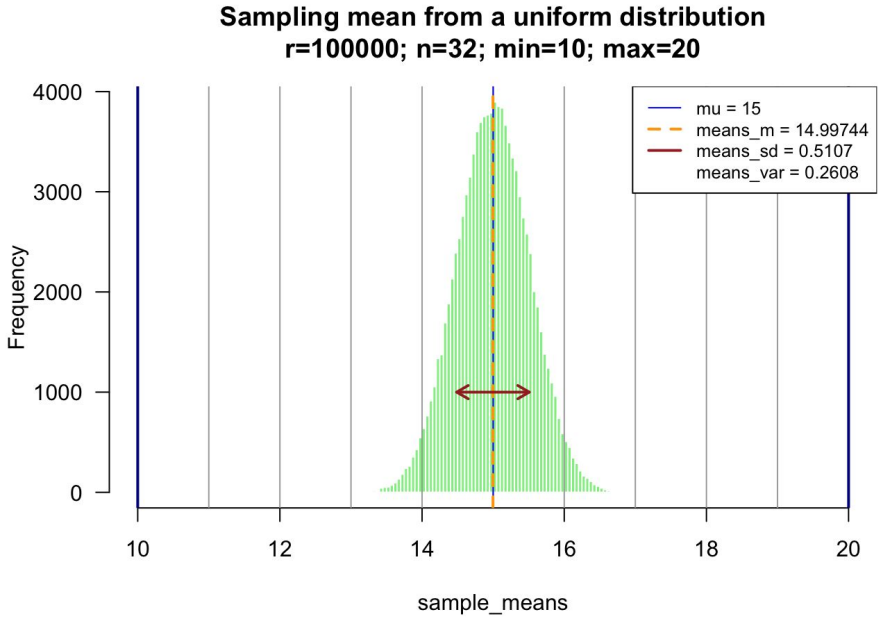
Distribution de la moyenne d'échantillons de n=16 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=16*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme en cloche, centrée sur 15 (la moyenne de la distribution uniforme).
- Note: la variance de ces moyennes de 16 nombres vaut 0.52.



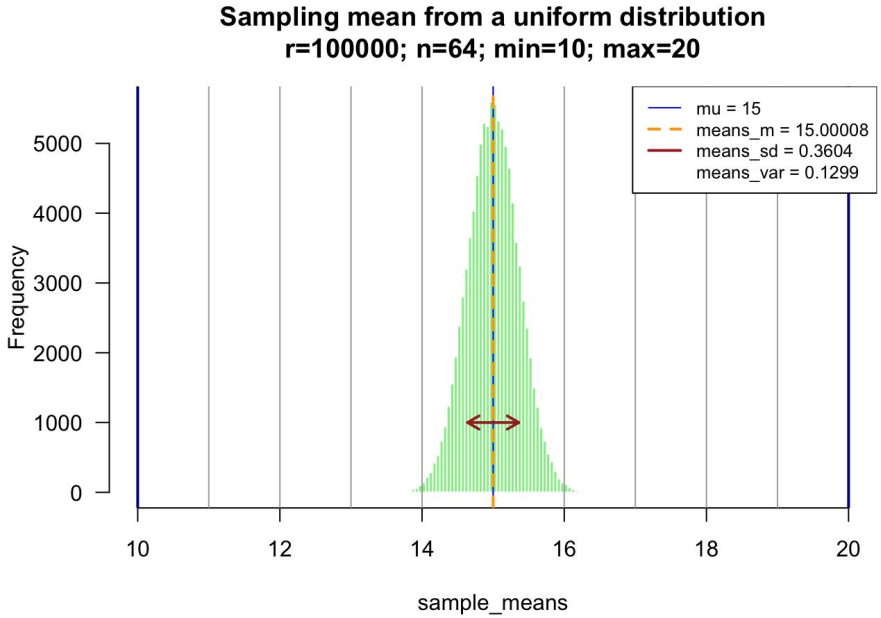
Distribution de la moyenne d'échantillons de n=32 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=32*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme en cloche, centrée sur 15 (la moyenne de la distribution uniforme).
- Note: la variance de ces moyennes de 32 nombres vaut 0.26.



Distribution de la moyenne d'échantillons de n=64 nombres

- L'histogramme montre la distribution des moyennes de *r=100.000* échantillons comportant chacun ***n=64*** nombres tirés dans une distribution uniforme s'étendant de +10 à +20.
- La distribution d'échantillonnage de la moyenne montre une forme en cloche, centrée sur 15 (la moyenne de la distribution uniforme).
- Note: la variance de ces moyennes de 64 nombres vaut 0.13.

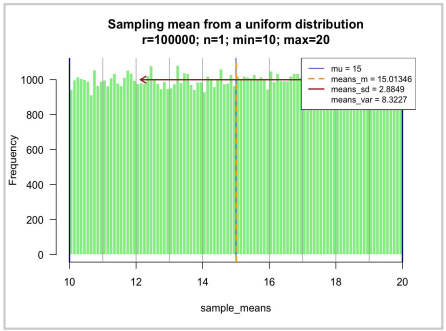


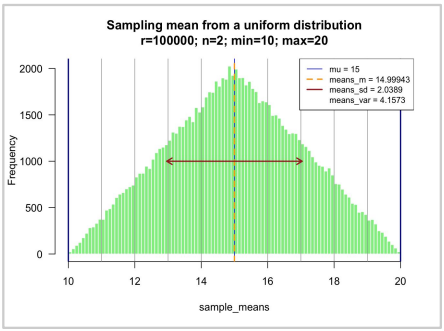
Impact de la taille d'échantillon sur la distribution d'échantillonnage de la moyenne

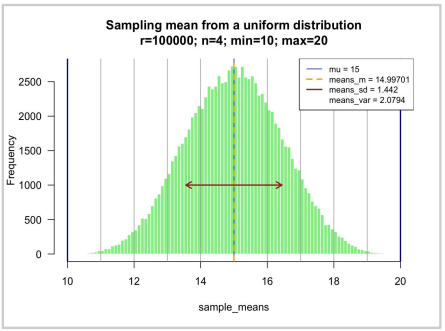
Quand la taille d'échantillon augmente, la distribution des moyennes d'échantillon

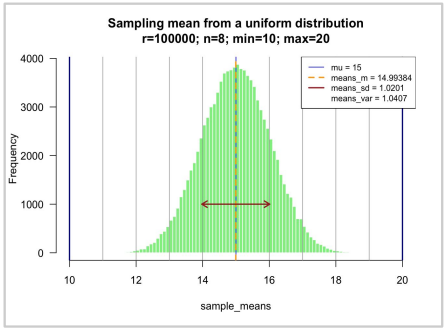
- Tend vers une normale (***théorème central limite***)
- Se resserre progressivement autour de la moyenne théorique
- La variance de la moyenne d'échantillon diminue linéairement en fonction de la taille de l'échantillon.
- L'écart type de la moyenne d'échantillon diminue *en raison de la racine de la taille d'échantillon*.

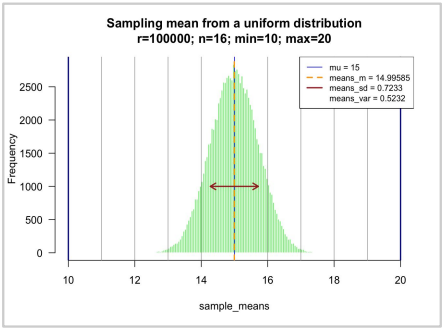
n	m	var	sd
1	15.01346	8.323	2.885
2	14.99943	4.157	2.039
4	14.99701	2.079	1.442
8	14.99384	1.041	1.020
16	14.99585	0.523	0.723
32	14.99744	0.261	0.511
64	15.00008	0.130	0.360

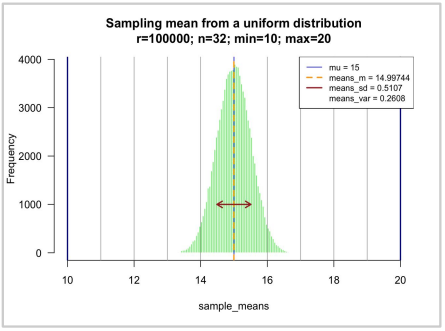


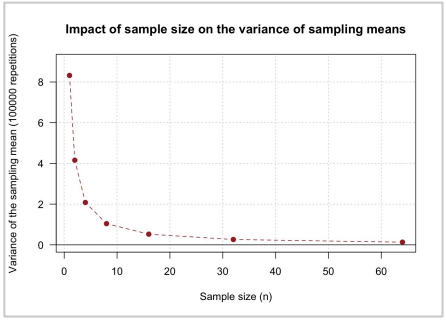


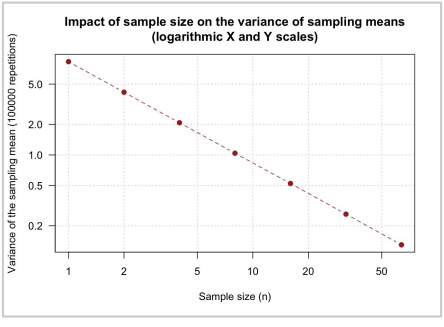


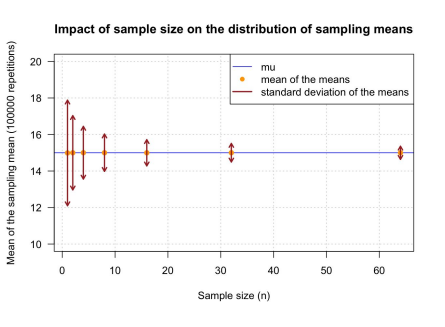












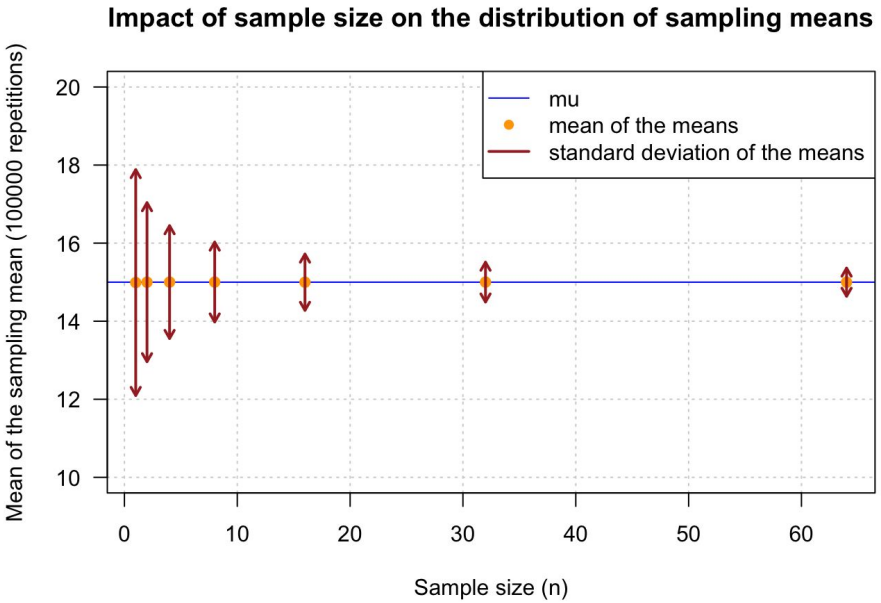
Erreur standard

On définit l'**erreur standard** comme l'écart-type de l'échantillonnage de la moyenne, autrement dit l'écart-type des moyennes de tous les échantillons qu'on pourrait tirer de la population.

- L'erreur standard égale l'écart-type de la population, divisé par la racine carrée de la taille de l'échantillon.
- On peut l'interpréter comme l'imprécision attendue sur la moyenne d'un échantillon.

Attention: en anglais, ne pas confondre

- **Standard error = erreur standard**
- **Standard deviation = écart-type**



Erreur standard

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{(n)}}$$

Sampling distribution of the mean

- In this simulation, the population is drawn randomly from a uniform distribution.
- When the sample size (*n*) increases, the sample mean tends towards a normal distribution. This is an application of the ***central limit theorem***.
- On the histograms of the previous slide, the distribution of the sample means is always centred around +15, irrespective of the sample size. The mean of the sample is an ***unbiased estimate*** of the population mean: its expected value equals the mean of the population.
- Note: the variance and standard deviation of the sample mean decrease as the sample size (*n*) increase.
- The expectation for the sample mean is the population mean. The sample mean is thus an ***unbiased*** estimator of the population mean.

$$E\left(\overline{X}\right) = m \qquad \hat{m} = \overline{X}$$

(the hat means "estimate")

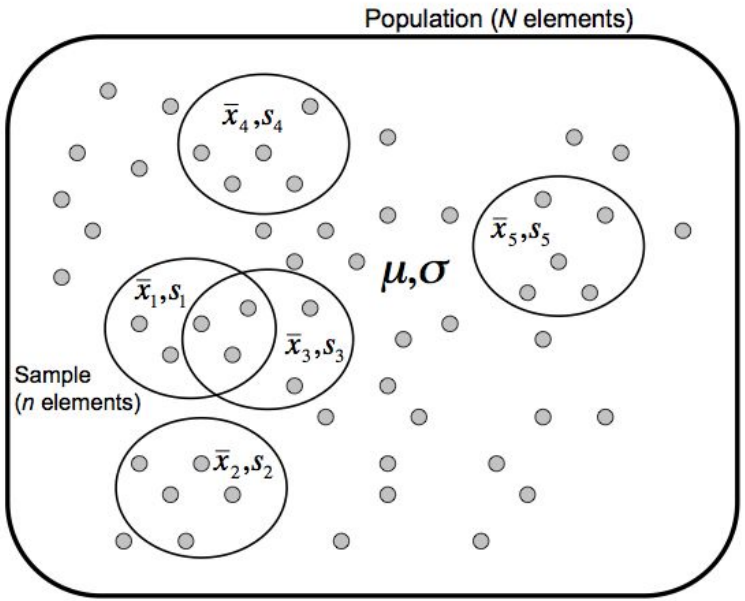
Sampling distribution - Sample variance

- The **sample variance** is a **biased** estimator of the population variance.
- Its expectation is lower than the actual variance $\sigma^2 \rightarrow$ the sample variance ***under-estimates*** the population variance.

$$E\left(S^2\right) = \frac{(n-1)}{n}\sigma^2 \qquad E(S) = \sqrt{\frac{(n-1)}{n}}\sigma$$

- To obtain an **unbiased estimation** of the population variance, one has to introduce a corrective factor ***n/(n-1)***.

$$\hat{\sigma}^2 = \frac{n}{n-1}s^2 \qquad \hat{\sigma} = \sqrt{\frac{n}{n-1}}s$$



Remarks

- This correction only matters for small samples : for very large samples, n/(n-1) ~ 1.
- This correction is already included in some packages.
  - In R, the function var() does not return the actual variance of the input vector (generally your sample), but a bias-corrected estimate for population variance.
  - The same holds true for the R sd() function.
  - **This can be misleading → beware of it**

Sampling distribution - The **standard error**

The expectation for the sample mean is the population mean.
The **sample mean** is thus an **unbiased** estimator of the **population mean**.

$$E\left(\overline{X}\right) = \mu \qquad \hat{\mu} = \overline{x}$$ *(the hat means "estimate")*

The variance of the sample mean distribution **differs** from the population variance.
The **sample variance** is thus a **biased** estimator of the **population variance** (this also applies to the standard deviation).

for a finite population

$$\sigma^2_{\overline{X}} = \frac{\sigma^2}{n}\left(\frac{N-n}{N-1}\right)$$

for an infinite population

$$\sigma^2_{\overline{X}} = \sigma^2 / n$$

The **standard deviation of the sample mean** is called ***standard error***. It decreases when *n* increases.
The larger is the sample, the more reliable is the estimation of the mean.

For a finite population

$$\sigma_{\overline{X}} = \frac{\sigma}{\sqrt{n}}\sqrt{\left(\frac{N-n}{N-1}\right)}$$

For an infinite population

$$\sigma_{\overline{X}} = \sigma / \sqrt{n} \qquad \hat{\sigma}_{\overline{X}} = \frac{s^2}{n-1}$$

Points-clés : échantillonnage et estimation

Échantillonnage

Tirage aléatoire d'un sous-ensemble d'éléments d'une population

Théorème central limite

La distribution d'échantillonnage de la moyenne tend vers une normale, indépendamment de la distribution de la population.

Estimation

- Tentative de caractériser des propriétés de la population à partir de celles d'un l'échantillon
- **Estimateurs**
  - La **moyenne d'échantillon** est un estimateur **non-biaisé** de la moyenne de population

  La **variance d'échantillon** est un estimateur **biaisé** de la variance de la population, mais on peut corriger le biais en divisant par *n-1* au lieu de *n*.

  Idem pour l'écart-type (mais en divisant par la racine de *n-1*)
- **Précision de l'estimation de la moyenne**
  - **L'erreur-standard** est l'écart-type de la moyenne d'échantillons.

  Sa précision augmente comme la racine carrée de la taille d'échantillon (*n*)

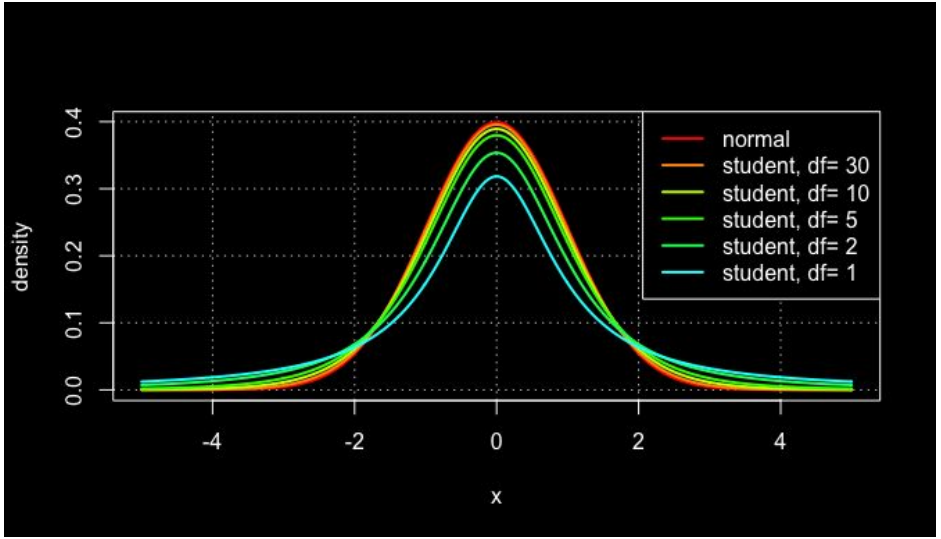
  → pour doubler la précision, il faut quadrupler la taille de l'échantillon.

Test de comparaison de moyennes

Détection des protéines dont l’abondance diffère entre foie et hépatocarcinome

Distribution de Student

- Une famille de courbes caractérisées par 1 paramètre : le nombre de **degrés de liberté** (df)
- La forme varie selon df
- Converge vers une normale quand df $\longrightarrow \infty$



Effet de la taille de l’échantillon sur l’intervalle de confiance

La précision de l’estimation augmente en raison de la racine de la taille d’échantillon. Pour doubler la précision, il faut donc quadrupler l’effectif de l’échantillon !

Exemple pratique : impact des valeurs manquantes sur les intervalles de confiance de l’abondance dans les données protéomique