Statistical Literacy — MINT

Lecture 6: All roads lead to the Normal distribution - The sampling distribution

Rémi Viné

The Graduate Institute | Geneva

October 30th, 2023

Outline

Housekeeping

The sampling distribution

The central limit theorem

Compute probabilities using the sampling distribution

Housekeeping

- Problem set 6 is now available
- You will be given assignment 1's grade on the week of the 20th of November
- Exam: some information
  - Closed choice, on Moodle
  - Calculator
  - Laptop

A first and fundamental step in inferential statistics

- Usually, we have the sample average ($\bar{x}$) and the sample standard deviation (s)
- From this, we want the have an idea on the population average (μ) and the population standard deviation (σ)

- For this, we need to introduce a key concept: the **sampling distribution**
  - This is a *theoretical concept*, it is unlikely to ever construct one
  - Here we focus on the *sampling distribution of sample averages*
    - But there can be as many sampling distributions as indicators (median, variance, *etc.*)

Construct a sampling distribution

Ingredients

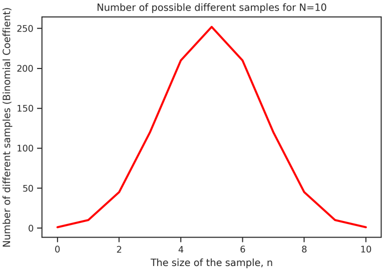
1. Choose a sample size ($n = \cdot$)
2. Pick one sample in the population and find the average ($\bar{x_1}$)
3. Store it
4. Choose another sample of size n, get $\bar{x_2}$
5. Continue this until all possible samples have been selected (you end up having m samples, and thus m sample averages)

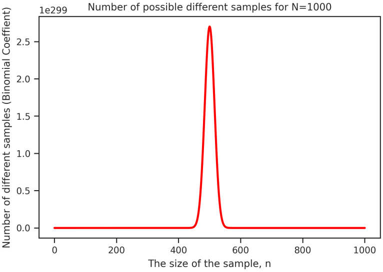
Sample 1	$\bar{x_1}$
Sample 2	$\bar{x_2}$
...	...
Sample m	$\bar{x_m}$

- The sequence of $\bar{x_i}$ is a random variable $\bar{X}$
  - Different samples $\rightarrow$ different statistics (average, Q2, *etc.*)

The number of samples becomes quickly gigantic

- Assume the population is 3 people (then statistics is probably useless...). There exist
  - 1 sample of $n = 3 \rightarrow$ ({1,2,3})
  - 3 samples of $n = 2 \rightarrow (\{1, 2\}, \{2, 3\}, \{1, 3\})$
  - 3 samples of $n = 1 \rightarrow$ ({1},{2},{3})
- But when the population increases, it becomes untraceable:

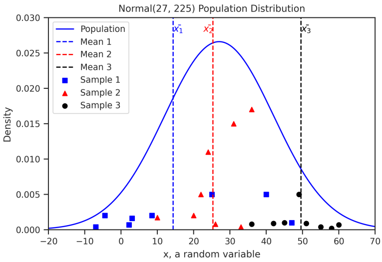




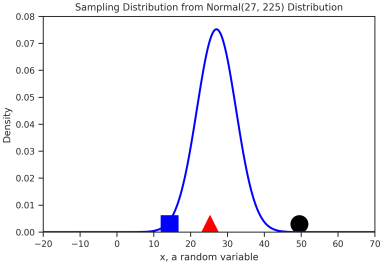
Construction of the sampling distribution

- When constructing sample averages, some averages are more likely
  - Here, $\bar{x_2}$ (or closely around) is more likely than $\bar{x_3}$ because more samples would end up with a sample average close to the population average

(a) From population distribution...

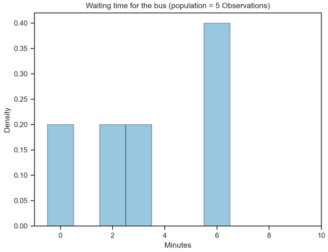


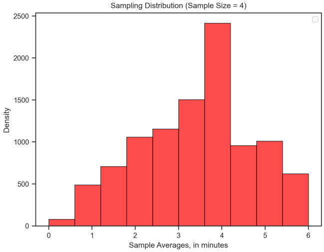
(b) ... to sampling distribution



Construction of the sampling distribution ($n = 4$)

- In the previous example, samples were picked from a population distribution that is Normally distributed
- But what if the population distribution is not Normal?





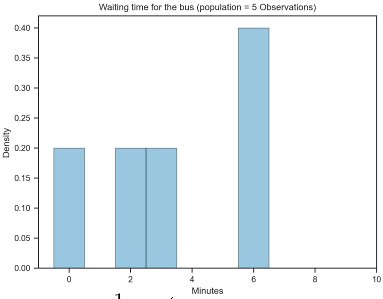
- Population distribution far from Normal
- Sampling distribution is already closer to a symmetric (though not Bell-Shaped yet)
  - The sampling distribution is more compact and more Bell-shaped (than the population distribution)

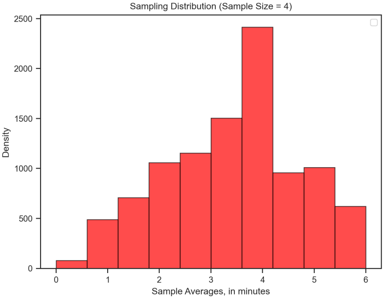
Properties of the Sampling distribution (for averages)

- The average of the sampling average: $\mu_{\bar{x}} = \mu$
  - Hence, population average and sampling distribution average are **identical**
- The standard deviation of the sampling average: $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
  - This is also called the **Standard Error**
    - If willing to understand the formula, see *extra note* (not required)
  - Hence, it is related to
    - The standard deviation of the population distribution *(positively related)*
    - The size of the samples *(negatively related)*
  - $for\ n > 1, \sigma_{\bar{x}} < \sigma$: the sampling distribution is more **"compact"** than the population distribution

Properties of the Sampling distribution (for averages)

With previous example (sample size, $n = 4$)





- $\mu = \frac{1}{5} \times (0 + 2 + 3 + 6 + 6) = 3.4\ minutes$
- $\mu_{\bar{x}} = 3.4\ minutes$ *(calculations would be a bit tedious, but visually this is - more or less - clear)*
- $\sigma = \sqrt{\frac{1}{5} \times (0 + 4 + 9 + 36 + 36) - 3.4^2} = \sqrt{5.44} \approx 2.33\ minutes$
- $\sigma_{\bar{x}} \approx \frac{2.33}{\sqrt{4}} \approx 1.17\ minutes$

The Central Limit Theorem

- This is perhaps the most important theorem in statistics
  - And it is particularly elegant

"I know of scarcely anything so apt to impress the imagination as the wonderful form of cosmic order expressed by the [Central Limit Theorem]. The law would have been personified by the Greeks and deified, if they had known it. It reigns with serenity and in complete self-effacement amidst the wildest confusion"

Sir Francis Galton, Natural Inheritance, p66.

The Central Limit Theorem

Statement (simplified)

- As the sample size, n, increases, the **sampling distribution** looks more and more like a less and less dispersed Normal distribution
- Given a population with mean μ and standard deviation σ
  - Taking random samples of size n large enough from the population of interest
    - Here, **"large enough" boils down to 30**
  - Then the distribution of sample means is approximately a Normal distribution such that: $\bar{X} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$

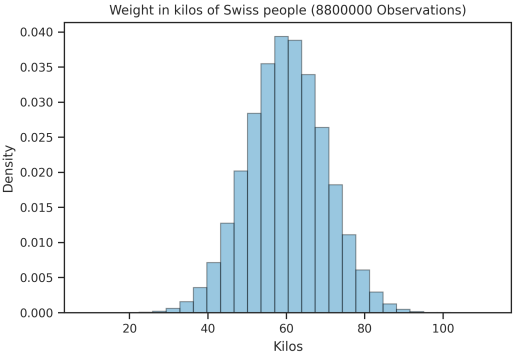
The Central Limit Theorem

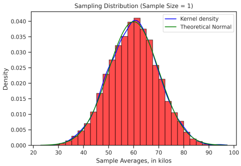
More details

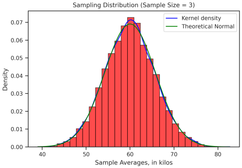
- Note that if the population distribution is Normal, then the sampling distribution is normal *regardless* of the sample size
- The less the population data looks like a bell-shaped distribution the bigger the sample size required
  - Here, the rule is (over)simplified to: "the sampling distribution is approximately Normal if the sample size is $n > 30$ (or if the population data is Normal)"

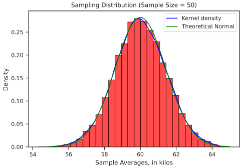
The Central Limit Theorem

With a population distribution close to Normal



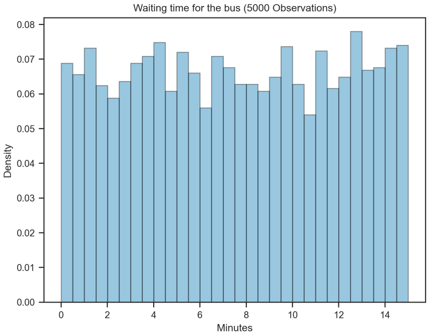


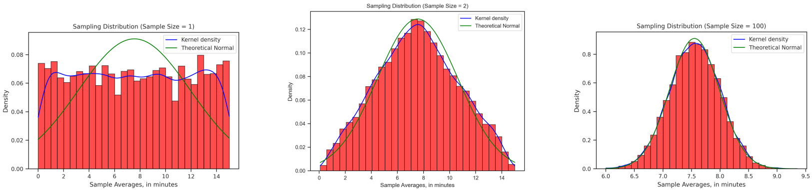




The Central Limit Theorem

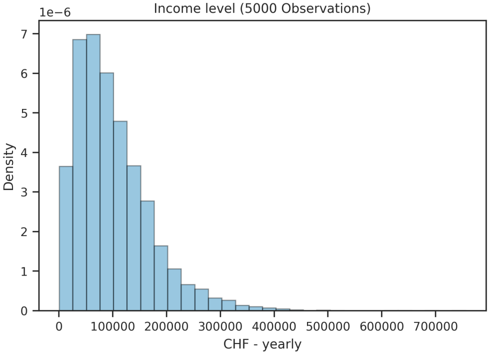
With a population distribution close to Uniform

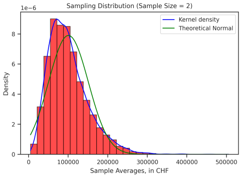


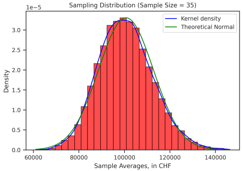


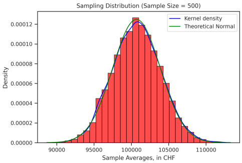
The Central Limit Theorem

With an asymmetric population distribution









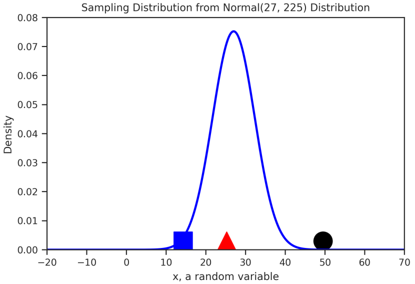
The Central Limit Theorem

The key to inference

- Inferences are based on the sampling distribution
- Knowing properties of this distribution is fundamental for the inference
  - One can get an idea about how likely is the current sample in use, *provided that the sample size is large enough*
  - Surveying an extremely rare sample may give an information about whether this sample is in fact different from the rest of the population (hopefully having an idea why)

Probability in the sampling distribution

- What is the probability to have a sample like the square, the triangle, the disk (of the figure below)?
- It seems clear that the sample with $\bar{x}$ far away from the sampling distribution mean ($\mu_{\bar{x}}$) is less likely
  - It is in a tail and implies to sample almost only observations with large values
- But concretely, how to calculate the probabilities?



Probability in the sampling distribution

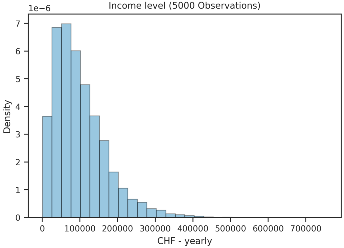
- Concretely, how to calculate the probabilities?
- Since the sampling distribution is Normal ($\bar{X} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$)
  - One simply needs to *standardize* and then find the corresponding probability to the *z-score* (or *vice versa*)
  - Standardizing the sampling distribution adds a slight refinement:

$$z = \frac{\bar{x} - \mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu_{\bar{x}}}{\frac{\sigma}{\sqrt{n}}}$$

- Remember: standardizing (finding the *z-score*) for the original distribution was done as $z = \frac{x-\mu}{\sigma}$
  - Hence, the idea is unchanged, we only adapts the centrality and dispersion indicators to the *new* distribution of interest

Probability in the sampling distribution

Example



- In a population of 5000 employees in the Humanitarian sector, the income distribution is the one displayed here (fake data)
  - The mean is $\mu \approx 100679$ in CHF
  - The standard deviation is $\sigma \approx 71354$ in CHF
- What is the probability to select a sample (randomly) of 81 employees whose average income is lower than $\bar{x} = CHF85600$?

Probability in the sampling distribution

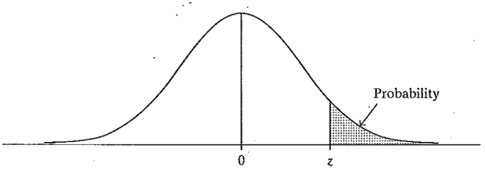
Example

- In a population of 5000 employees in the Humanitarian sector, the income distribution is the one displayed here (fake data)
  - The mean is $\mu \approx 100679$ in CHF
  - The standard deviation is $\sigma \approx 71354$ in CHF
- What is the probability to select a sample (randomly) of 81 employees whose average income is $\bar{X} < 85600$ Swiss Francs?
- $P(\bar{X} < 85600) = P(Z < \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}})$
  - Compute the standard error: $\sigma_{\bar{x}= \frac{\sigma}{\sqrt{n}} = \frac{71354}{\sqrt{81}}} \approx 7928$ CHF
- $P(\bar{X} < 85600) = P(Z < \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}) = P(Z < \frac{85600-100679}{7928}) = P(Z < -1.90) = P(Z > 1.90)$

Probability in the sampling distribution

Example

TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z, probabilities are found by symmetry)



z	Second Decimal Place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
...										
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084

Probability in the sampling distribution

Example

- In a population of 5000 employees in the Humanitarian sector, the income distribution is the one displayed here (fake data)
  - The mean is $\mu \approx 100679$ in CHF
  - The standard deviation is $\sigma \approx 71354$ in CHF
- What is the probability to select a sample (randomly) of 81 employees whose average income is $\bar{x} < 85600$ Swiss Francs?
- $P(\bar{X} < 85600) = P(Z < \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}) = P(Z < \frac{85600-100679}{7928}) = P(Z < -1.90) = P(Z > 1.90) = 2.87\%$
- **Selecting a random sample of 81 employees where the average wage is lower than CHF 85600 can happen with a probability 2.87%**

Probability in the sampling distribution

Example

- What if, instead of 81 employees, only 18 were sampled?
- What is, instead, the sample was of 361 employees?

Probability in the sampling distribution

Example, cont'd

- What if, instead of 81 employees, only 18 were sampled?
  - Here the sampling distribution cannot be considered as Normal and thus we cannot compute probabilities (in the class)
- What is, instead, the sample was of 361 employees?

Quick wrap-up on probabilities

Original population X	Sampling distribution $\bar{X}$
Any types of distribution (μ, σ)	Normal distribution for $n > 30$, $\mu_{\bar{x}} = \mu$ & $\sigma_{\bar{x}} = \sigma$
$P(an\ observation\ in\ an\ interval)$	$P(a\ sample\ in\ an\ interval)$
z-score for single observation	*z-score* for sample
$z = \frac{x-\mu}{\sigma}$	$z = \frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}} = \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}}$

Next session

- Next session is on confidence intervals (a first inference tool)