Lecture 4 - sampling

Dr. L.S. Stephan

2023

RUG Groningen

learning objectives

you will learn about

1. how the population and the sample are linked through the concept of convergence
2. why estimators have a variance and a distribution
3. why larger samples are a remedy for many, but not all, of our problems

convergence - the link between the sample and the population

- introduction
- the law of large numbers (LLN)

estimator variance - why we generally prefer large samples

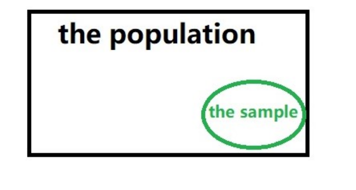
- introduction
- the central limit theorem CLT
- the importance of N
- trade-offs - why do we sometimes NOT prefer a large sample size

convergence - the link between the sample and the population

convergence - the link between the sample and the population

introduction

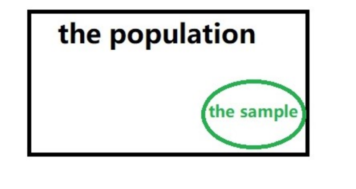
recap- sample and population



frequency: $f_j = \frac{\sum_{i=1}^{N} \mathbf{I}(x_i = j)}{N}$ fraction of category j in the **sample**

probability: $P(X_i = j)$ fraction of category j in the **population**

recap- sample and population

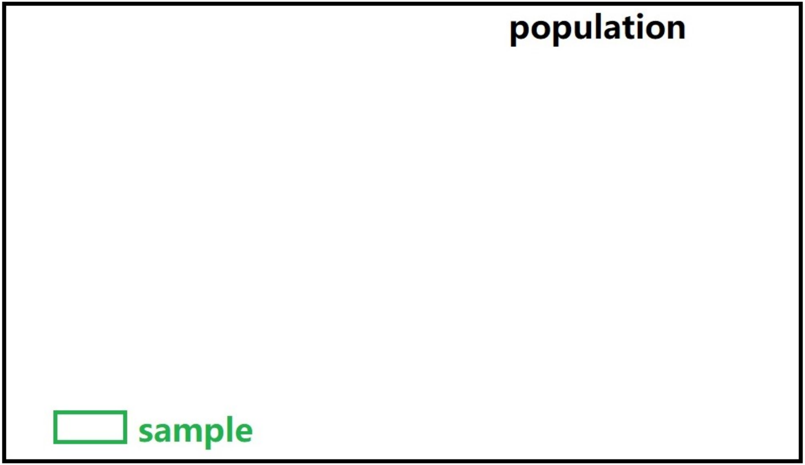


frequency: $f_j = \frac{\sum_{i=1}^{N} \mathbf{I}(x_i = j)}{N}$ fraction of category j in the **sample**

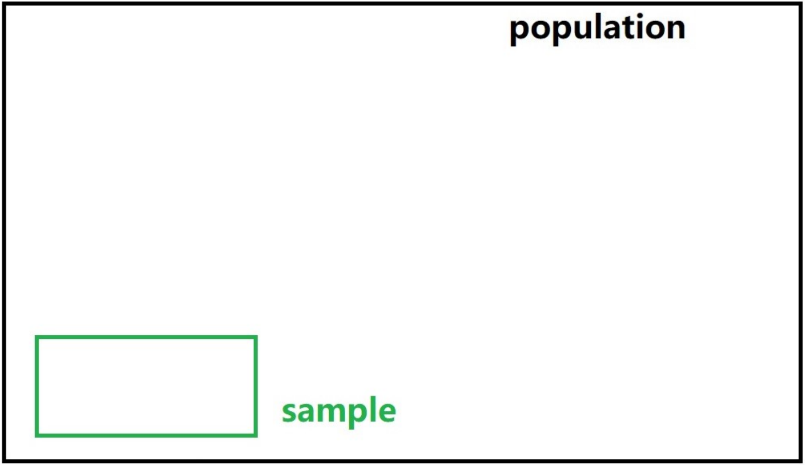
probability: $P(X_i = j)$ fraction of category j in the **population**

what happens when we make our sample larger and larger ?

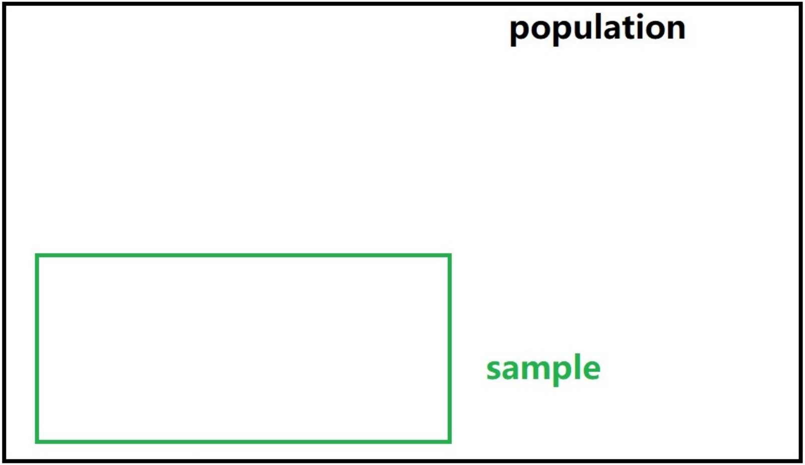
larger samples



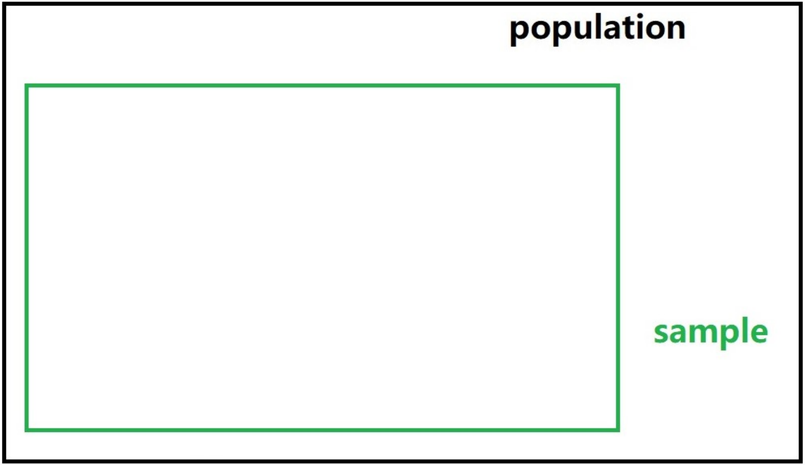
larger samples



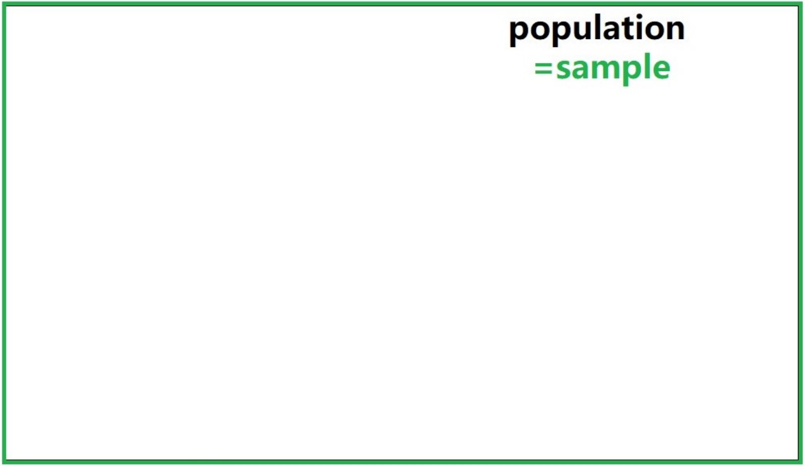
larger samples



larger samples



larger samples



population

=sample

Frame Title

- (given our sample is representative), the sample will resemble the population more and more
- remember that we characterise both sample and population by **moments**
- we also characterise the sample by **frequencies** and the population by **probabilities**
- "becoming more similar" means that the distance between sample and population moments shrinks, as well as the distance between frequencies and probabilities
- we call this **convergence**
- the field of study that investigates the behaviour of estimates as the sample size goes to infinity (get VERY large) is called **asymptotics**

convergence - the link between the sample and the population

the law of large numbers (LLN)

the law of large number

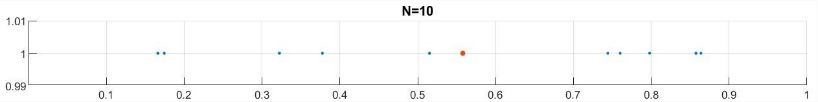
if we sample N observations independently at random from the same population, then the LLN states that as N grows VERY large, *"sample moments converge in probability to the corresponding population moments. In other words, the probability that the sample mean is close to the population mean can be made as high as you like by taking a large enough sample."* (Mostly Harmless Econometrics by J. D. Angrist and J. S. Pischke)

LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ $(\Rightarrow \mu_X = 0.5)$

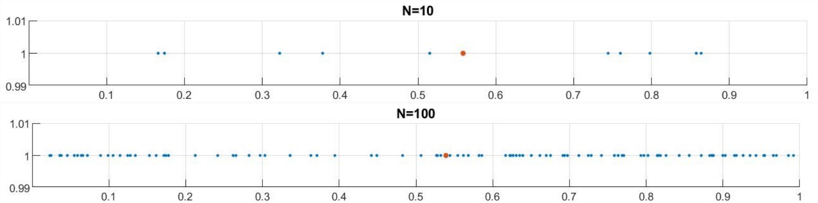
LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ $(\Rightarrow \mu_X = 0.5)$



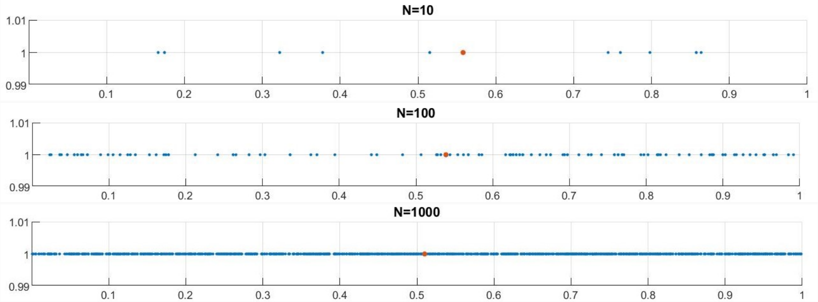
LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ $(\Rightarrow \mu_X = 0.5)$



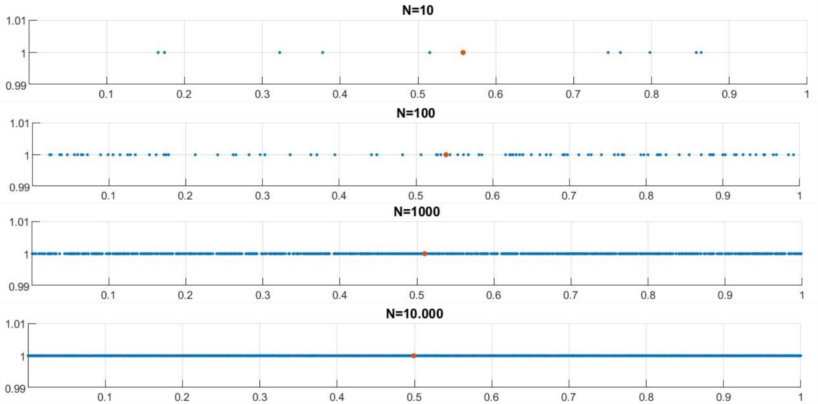
LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ ($\Rightarrow \mu_X = 0.5$)



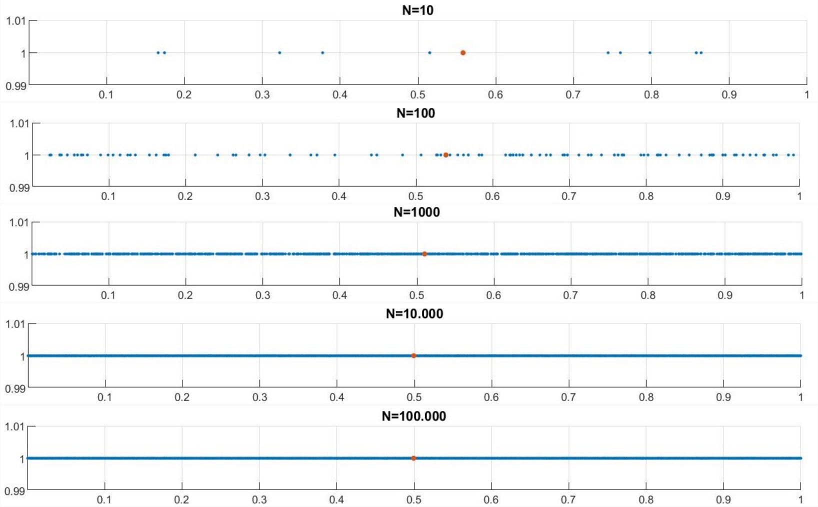
LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ $(\Rightarrow \mu_X = 0.5)$



LLN: *"...In other words, the probability that the* **sample mean** *is close to the* **population mean** *can be made as high as you like by taking a* **large enough** *sample."*

let us verify this statement! we draw at random from $X \sim U(0,1)$ ($\Rightarrow \mu_X = 0.5$)



we digest the law of large numbers (LLN)

- the characteristic we study is distributed according to some probability distribution in the population (we say $X \sim f_x$, in the example above $X \sim U(0,1)$)
- then if we randomly draw an individual (i) and measure the characteristic, this is a **random experiment**
- consequently, (ex ante) the characteristics of a randomly drawn individual is a random variable (X_i)
- after the experiment is performed, we know x_i, the realisation of the random variable X_i, namely, the characteristic of the person we have drawn
- if we intend to repeat this N times, we consider a sequence of N random variables $X_1, ..., X_N$
- after we performed the experiment, we have one realisation for each of these random variables

we digest the law of large numbers (LLN)

- we consider the sequence of random variables $X_1, ..., X_N$
- because the individuals are drawn **independently at random** from the **same population**, they follow the **same probability distribution** (that of the population)
- we say they are i.i.d. (independent and identically distributed)
- now we consider the sample mean $\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$
- this is function of random variables, hence also a random variable
- for our particular sample, the realisation of the sample mean will be $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$, it will be different for each sample
- the LLN states that, if N "goes to infinity" the sample mean essentially ceases to be random: **it will coincide with the population mean for sure!**

we digest the law of large numbers (LLN)

if we sample N observations independently at random from the same population, that means then the LLN states that as N grows VERY large, *"sample moments converge in probability to the corresponding population moments. "*

$$\lim_{N\to\infty} P\left(|\bar{X}-\mu_x| \geq \varepsilon\right) = 0 \quad \forall \varepsilon > 0$$

the probability that the difference between the sample mean and the population mean exceeds an arbitrarily small number is zero when N is large enough and we say that

$$\bar{X} = \sum_{i=1}^{N} X_i \xrightarrow{p} \mu_x$$

"the sample mean converges in probability to the population mean"

$\Rightarrow$ if N is large enough, $\bar{x} = \bar{X} = \mu_x$

we apply the LLN - what about frequencies????

Lecture 1: an indicator function takes the value one if the condition is true and the value zero otherwise $\mathbf{I}(x_i = j) = 1$ if $x_i = j$

Lecture 1: a frequency $f_j = \frac{\sum_i \mathbf{I}(x_i=j)}{N}$ is a function of the sample observations

lecture 2: the probabilities are the frequencies for the population

now we understand why, namely, frequencies are realisations of a sample means!

- each $\mathbf{I}(x_i = j)$ is a realisation of the random variable $\mathbf{I}(X_i = j)$
- these random variables are i.i.d.
- $f_j = \frac{\sum_i \mathbf{I}(x_i=j)}{N}$ is a realisation of $\frac{\sum_i \mathbf{I}(X_i=j)}{N}$
- $\frac{\sum_i \mathbf{I}(X_i=j)}{N}$ is a sample mean of i.i.d. random variable
- $\Rightarrow$ the LLN applies

we apply the LLN - what about frequencies????

frequencies are realisations of sample means (of indicator functions) ⇒ means of realisations of i.i.d. random variable ⇒ the LLN applies

$$\frac{\sum_i \mathbf{I}(X_i = j)}{N} \xrightarrow{p} = E\left(\frac{\sum_i \mathbf{I}(X_i = j)}{N}\right) = \frac{\sum_i E(\mathbf{I}(X_i = j))}{N} = \frac{Np_j}{N} = p_j$$

so we conclude that for large enough N we have

$$f_j = \frac{\sum_i \mathbf{I}(X_i = j)}{N} = p_j$$

the relative frequency f_j in the sample converge to the probability that the random variable, representing a randomly selected member of the population, exhibits characteristic j

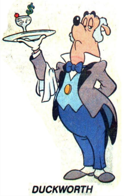
the population (of duckburg)

IQ:125,
wealth:100

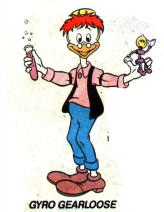
IQ:125,
wealth:60

IQ:100,
wealth:0

IQ:75
wealth:40



IQ:75,
wealth:40

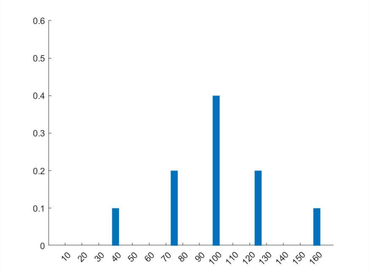


IQ: 160,
wealth:50

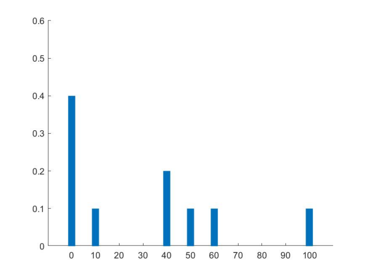
IQ: 100,
wealth:0

IQ: 40,
wealth:10

duckburg population probability distributions

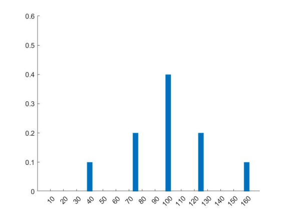


duckburg IQ probability distribution

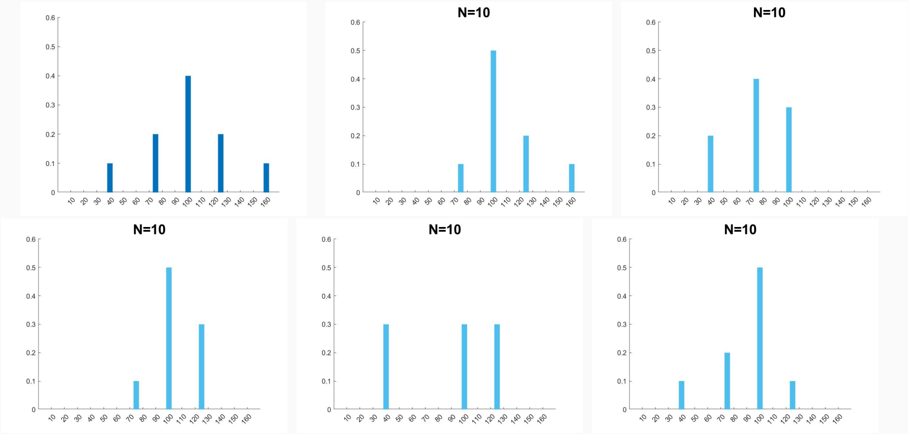


duckburg wealth probability distribution

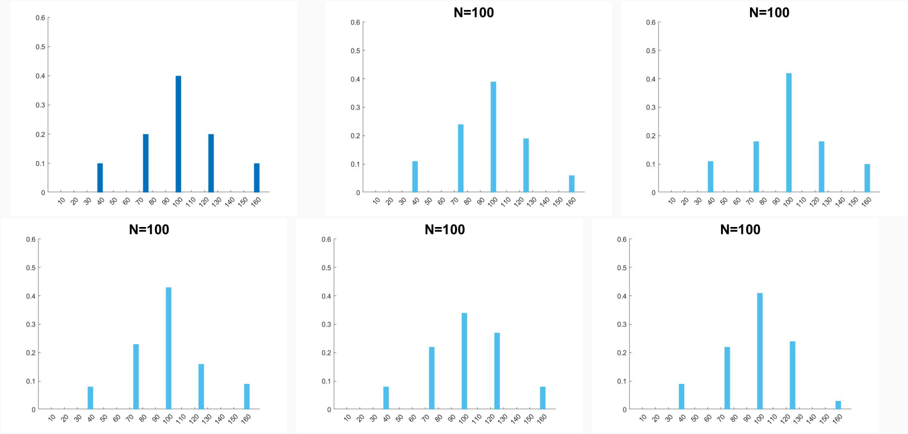
duckburg IQ population (N=1.000.000.000) v. sample



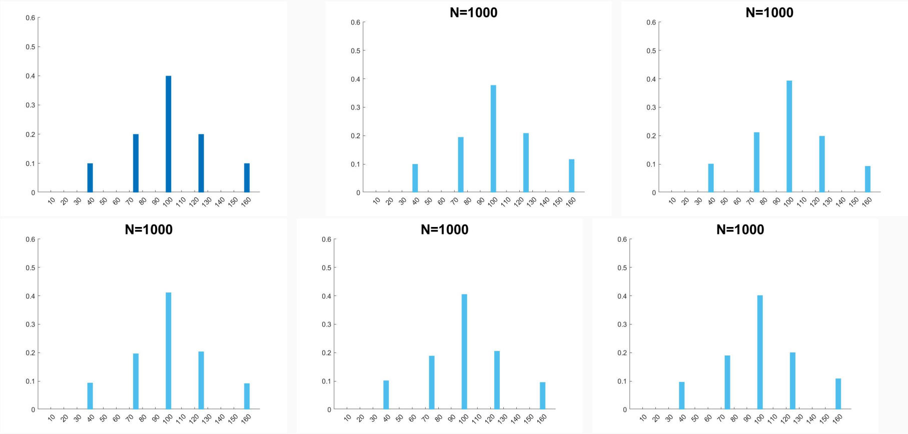
duckburg IQ population (N=1.000.000.000) v. sample



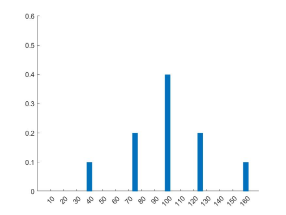
duckburg IQ population (N=1.000.000.000) v. sample

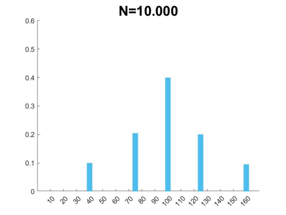


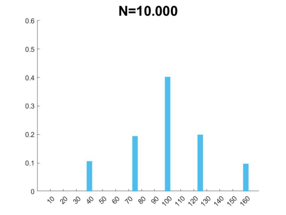
duckburg IQ population (N=1.000.000.000) v. sample

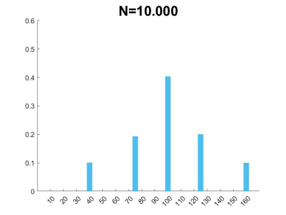


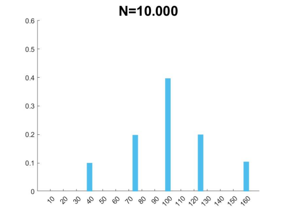
duckburg IQ population (N=1.000.000.000) v. sample

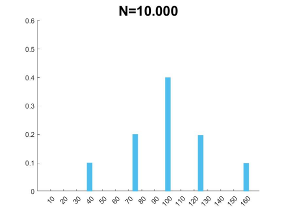



N=10.000

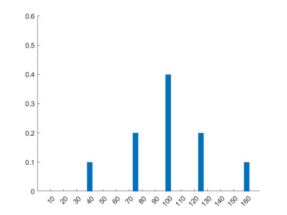

N=10.000

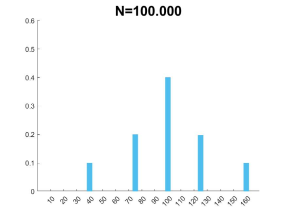

N=10.000

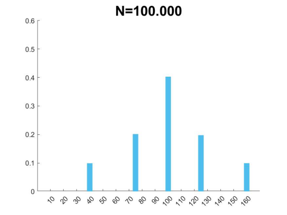

N=10.000

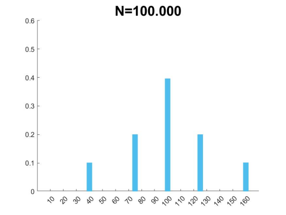

N=10.000

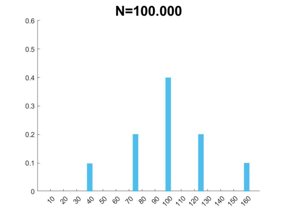
duckburg IQ population (N=1.000.000.000) v. sample

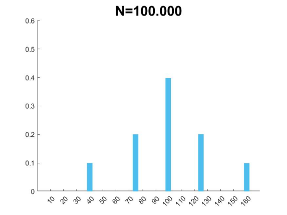



N=100.000

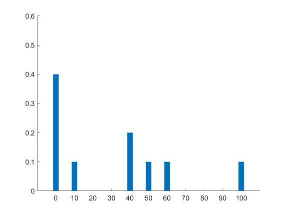

N=100.000


N=100.000

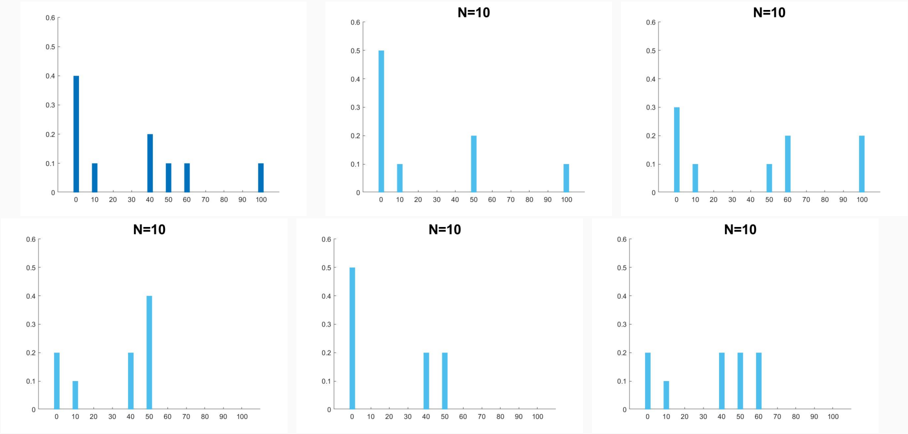

N=100.000


N=100.000

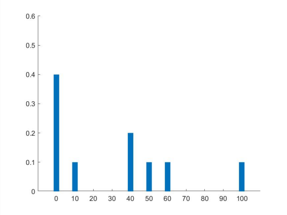
duckburg wealth population (N=1.000.000.000) v. sample

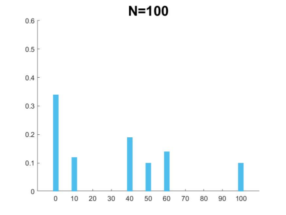


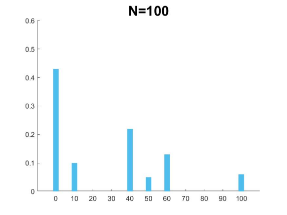
duckburg wealth population (N=1.000.000.000) v. sample

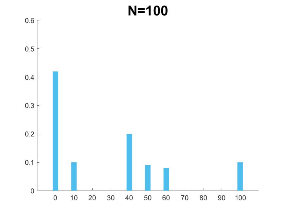


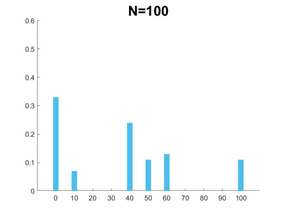
duckburg wealth population (N=1.000.000.000) v. sample

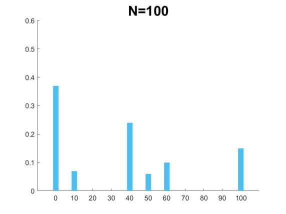




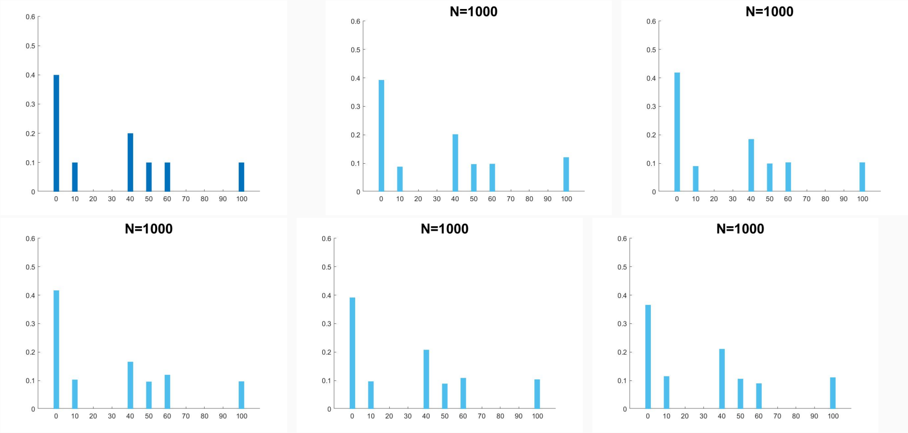




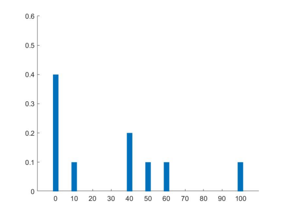


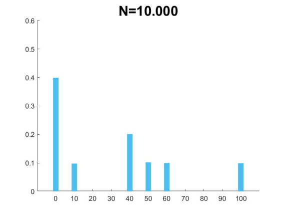


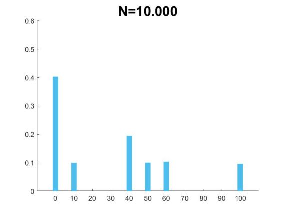
duckburg wealth population (N=1.000.000.000) v. sample

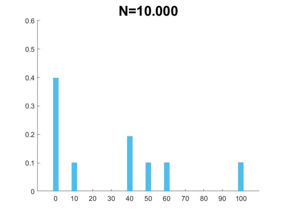


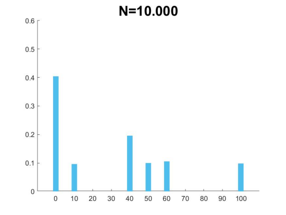
duckburg wealth population (N=1.000.000.000) v. sample

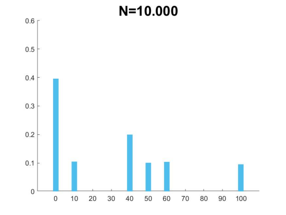



N=10.000

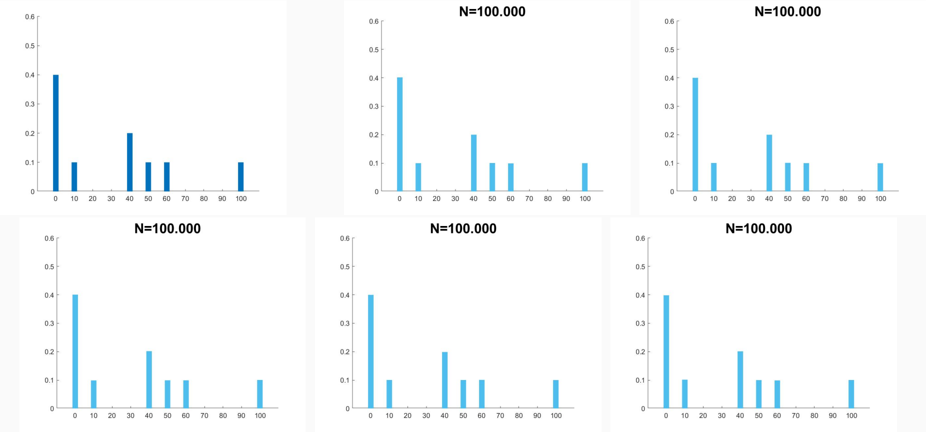

N=10.000


N=10.000


N=10.000


N=10.000

duckburg wealth population (N=1.000.000.000) v. sample



estimator variance - why we generally prefer large samples

estimator variance - why we generally prefer large samples

introduction

recap

we have learnt that

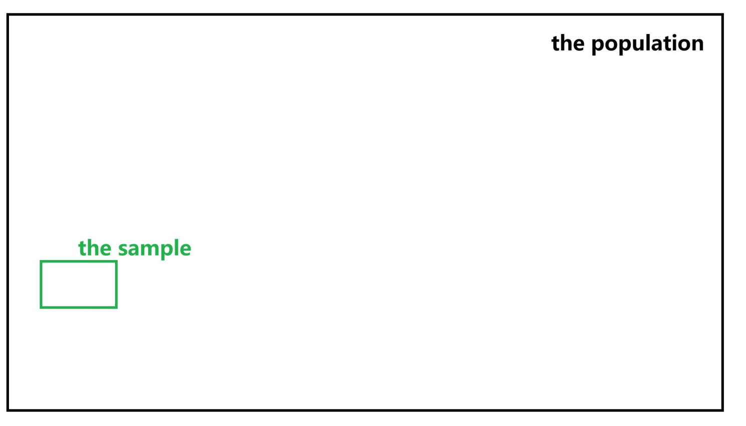
- if we sample N observations independently at random from the same population, the sample mean is random variable, with a different realisation each time we draw a sample
- if the sample is **large enough** and the sample is **representative**, the sample mean is not random, it coincides with the population mean

the sample mean is a random variable

... with a different realisation each time we draw a sample

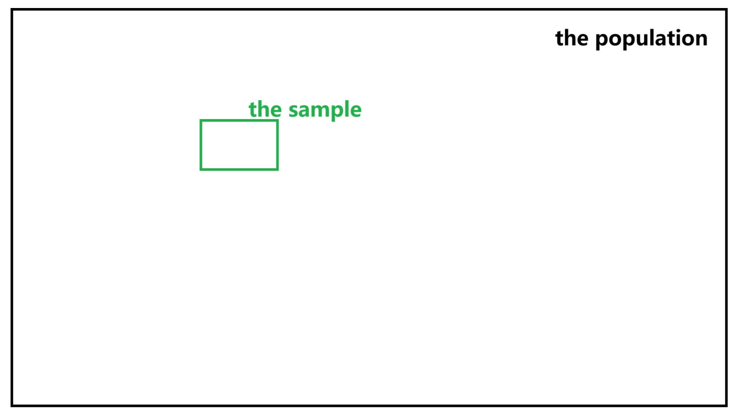
the sample mean is a random variable

… with a different realisation each time we draw a sample



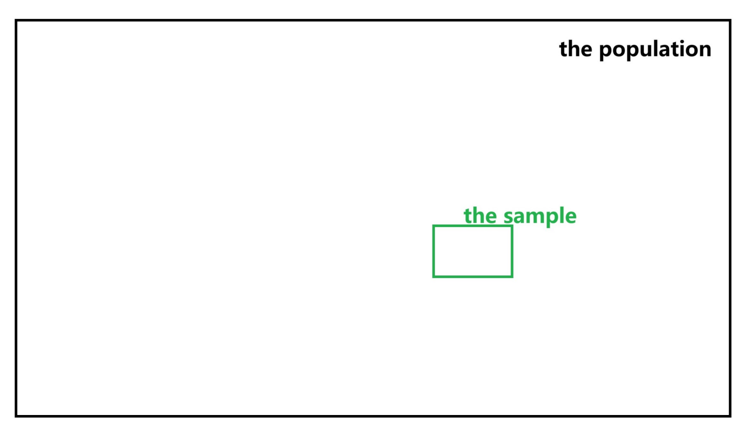
the sample mean is a random variable

… with a different realisation each time we draw a sample



the sample mean is a random variable

… with a different realisation each time we draw a sample



the sample mean is a random variable

$\Rightarrow$ it has a distribution!

- if we draw **more** samples, we obtain more estimate
- then by the LLN, the frequencies of those estimates converge to the true probabilities of the sample mean

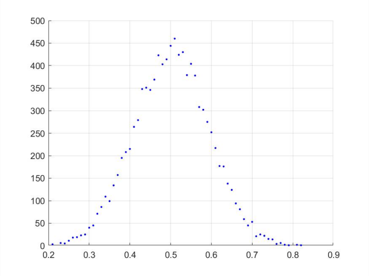
estimator variance - why we generally prefer large samples

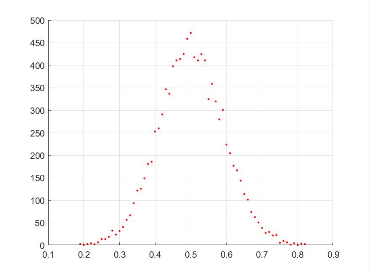
the central limit theorem CLT

illustration

again $X \sim U(0,1)$, **N=10** we draw 10.000 samples and compute the sample mean and as a comparison, we draw 10.000 times at random from a normal distribution:

here are the frequencies:



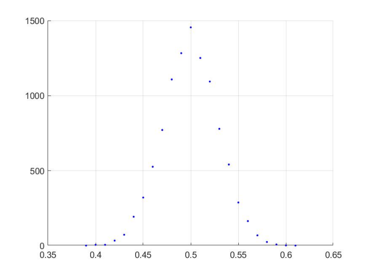


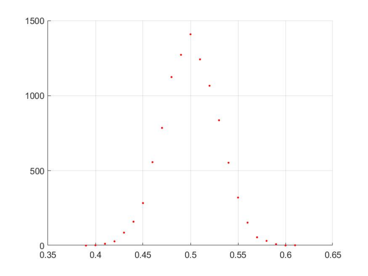
it looks so similar!!

illustration

again $X \sim U(0,1)$, **N=100** we draw 10.000 samples and compute the sample mean and as a comparison, we draw 10.000 times at random from a normal distribution:

here are the frequencies:



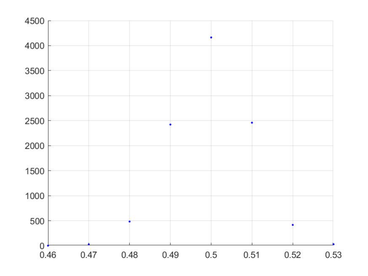


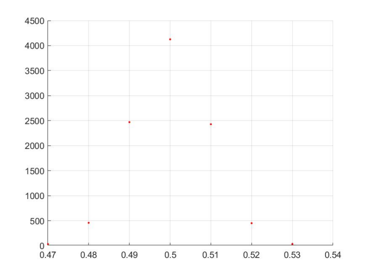
it looks so similar!!

illustration

again $X \sim U(0,1)$, **N=1000** we draw 10.000 samples and compute the sample mean and as a comparison, we draw 10.000 times at random from a normal distribution:

here are the frequencies:





it looks so similar!!

the central limit theorem CLT

this similarity in not a coincidence, the CLT states that *"sample moments are asymptotically Normally distributed*

mean and variance are those of the population, scales by the sample size

"In other words, in large enough samples, appropriately normalized sample moments are approximately standard Normally distributed" (Mostly Harmless Econometrics by J. D. Angrist and J. S. Pischke)

mean and variance of the sample mean

$$X \sim f_x(\mu_x, \sigma_x^2)$$

mean and variance of the sample mean

$$X \sim f_x(\mu_x, \sigma_x^2)$$

$$\bar{X} = \frac{1}{N}\sum_i X_i \quad \Rightarrow \quad \mu_{\bar{X}} = E\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N}\sum_i E(X_i) = \frac{N}{N}\mu_x = \mu_x$$

$$\sigma_{\bar{X}}^2 = E\left(\left(\frac{1}{N}\sum_i X_i - \mu_x\right)^2\right) = \frac{1}{N^2}\sum_i E(X_i - \mu_x)^2 = \frac{\sigma_x}{N}$$

mean and variance of the sample mean

$$X \sim f_x(\mu_x, \sigma_x^2)$$

$$\bar{X} = \frac{1}{N}\sum_i X_i \quad \Rightarrow \quad \mu_{\bar{X}} = E\left(\frac{1}{N}\sum_i X_i\right) = \frac{1}{N}\sum_i E(X_i) = \frac{N}{N}\mu_x = \mu_x$$

$$\sigma_{\bar{X}}^2 = E\left(\left(\frac{1}{N}\sum_i X_i - \mu_x\right)^2\right) = \frac{1}{N^2}\sum_i E(X_i - \mu_x)^2 = \frac{\sigma_x}{N}$$

$$\Rightarrow \bar{X} \sim N(\mu_x, \tfrac{\sigma_x^2}{N}) \Rightarrow \sqrt{N}(\bar{X} - \mu_x) \sim N(0, 1)$$

estimator variance - why we generally prefer large samples

the importance of N

we prefer large samples

- recap: the variance of the mean shrinks with N
- at some point, the variance is zero (remember: LLN)
- the mean converges
- this holds true for any estimator that is computed as a sample mean! (the majority of estimators you will encounter)
- if an estimator **converges** to its expected value (the variance goes to zero with the sample size) we say it it is **consistent**
- the expected value coincides with the truth, we say the estimator is **unbiased**
- for instance, the sample mean is a consistent and unbiased estimator of the population mean if
    - we sample independently at random
    - the sample is representative

the effect of the sample size, example 1

Do Elevated Viewpoints Increase Risk Taking?

(Journal of Marketing Research)

- $N = 203$
- significant effect

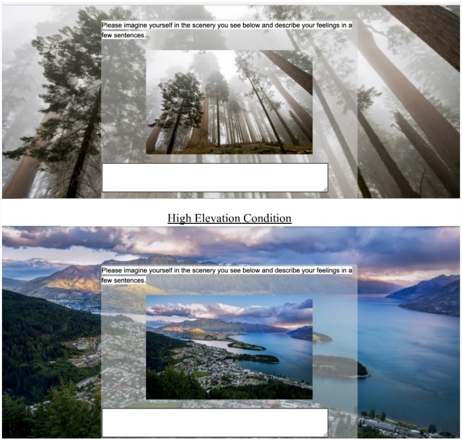
“people exposed to high-elevation sceneries [were] more willing to purchase new products than people exposed to low-elevation images”

replication study: https://datacolada.org

- $N = 203$
- **no** significant effect

two ostensibly separate studies

study one: rate pictures from either low or high position



Please imagine yourself in the scenery you see below and describe your feelings in a few sentences.

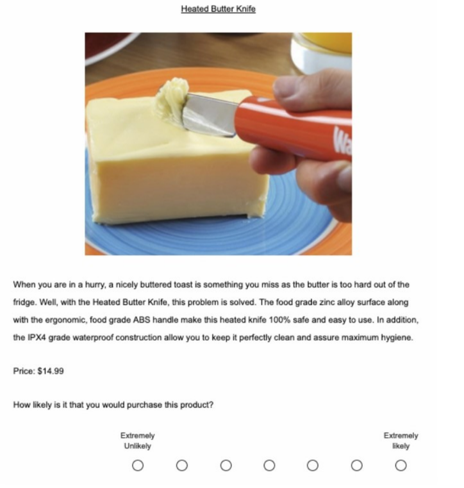
High Elevation Condition

Please imagine yourself in the scenery you see below and describe your feelings in a few sentences.

two ostensibly separate studies

study one: rate pictures from either low or high position

study two: rate likelihood to purchase novel product



Heated Butter Knife

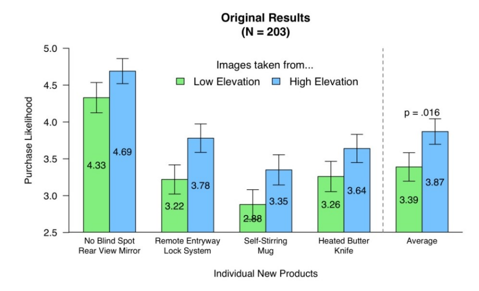
When you are in a hurry, a nicely buttered toast is something you miss as the butter is too hard out of the fridge. Well, with the Heated Butter Knife, this problem is solved. The food grade zinc alloy surface along with the ergonomic, food grade ABS handle make this heated knife 100% safe and easy to use. In addition, the IPX4 grade waterproof construction allow you to keep it perfectly clean and assure maximum hygiene.

Price: $14.99

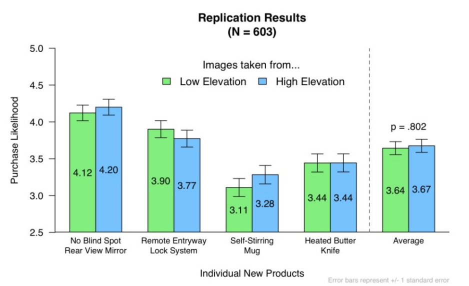
How likely is it that you would purchase this product?

Extremely Unlikely ○ ○ ○ ○ ○ ○ ○ Extremely likely

results



Original Results
(N = 203)
Images taken from...
🟩 Low Elevation
🟦 High Elevation
Purchase Likelihood
5.0
4.5
4.0
3.5
3.0
2.5
4.33
4.69
3.22
3.78
2.88
3.35
3.26
3.64
p = .016
3.39
3.87
No Blind Spot Rear View Mirror
Remote Entryway Lock System
Self-Stirring Mug
Heated Butter Knife
Average
Individual New Products



Replication Results
(N = 603)
Images taken from...
🟩 Low Elevation
🟦 High Elevation
Purchase Likelihood
5.0
4.5
4.0
3.5
3.0
2.5
4.12
4.20
3.90
3.77
3.11
3.28
3.44
3.44
p = .802
3.64
3.67
No Blind Spot Rear View Mirror
Remote Entryway Lock System
Self-Stirring Mug
Heated Butter Knife
Average
Individual New Products
Error bars represent +/- 1 standard error

the effect of the sample size, example 2

Does Displaying Multiple Copies of a Product Increase Its Perceived Effectiveness?

Journal of Consumer Research (JCR)

- $N = 87$
- significant effect

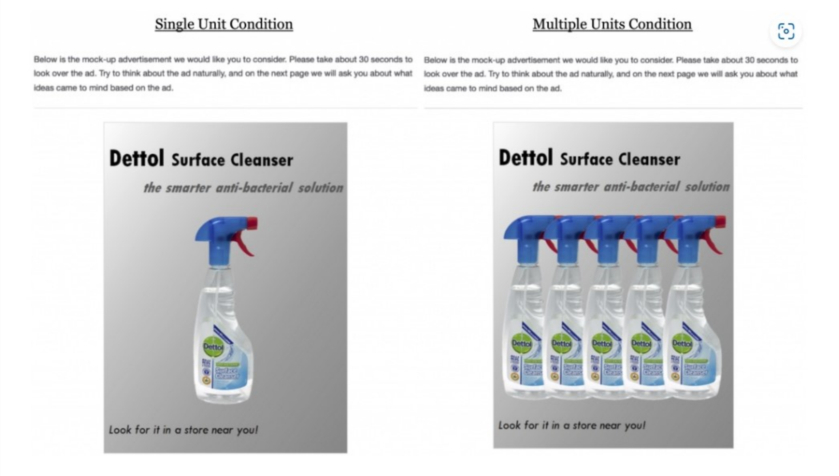
"participants who saw the ad displaying five bottles of Dettol judged it to be more effective than those who saw the ad displaying one bottle of Dettol"

replication study: https://datacolada.org

- $N = 636, N = 574, N = 702$
- **no** significant effect

the effect of the sample size - example 2

participants see one of the two adds and rate the perceived product efficiency



Single Unit Condition

Below is the mock-up advertisement we would like you to consider. Please take about 30 seconds to look over the ad. Try to think about the ad naturally, and on the next page we will ask you about what ideas came to mind based on the ad.

Dettol Surface Cleanser

the smarter anti-bacterial solution

Look for it in a store near you!

Multiple Units Condition

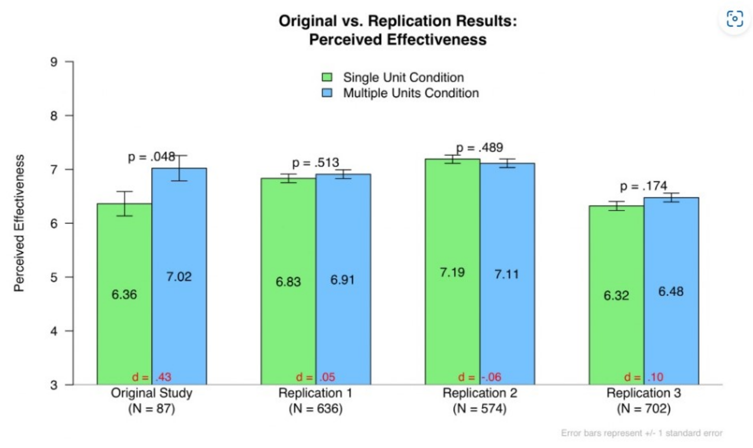
Below is the mock-up advertisement we would like you to consider. Please take about 30 seconds to look over the ad. Try to think about the ad naturally, and on the next page we will ask you about what ideas came to mind based on the ad.

Dettol Surface Cleanser

the smarter anti-bacterial solution

Look for it in a store near you!

the effect of the sample size - example 2



estimator variance - why we generally prefer large samples

trade-offs - why do we sometimes NOT prefer a large sample size

lesson learnt:

larger sample size ⇒ smaller variance ⇒ we can reject false hypothesis more easily!

Do we **always** want to sample more???

No! here is why:

- sampling is **costly**
- sample more ⇒ need more **time**
  - we may be under time pressure (competition!)
  - the DGP could change in the meantime

sampling is costly

- some samples are very costly to accumulate
- example: twin studies
- in general, hard to sample individuals that are rare
- increasing N may not be **feasible**

sampling is costly

- some samples are very costly to accumulate
- example: twin studies
- in general, hard to sample individuals that are rare
- increasing N may not be **feasible**

- some experiments raise ethical question (e.g. placebo tests)
- bargain: population benefits (research) versus individual cost
- we may not increase N for **ethical** reasons

sampling needs time!

you conduct a marketing survey that will guide the development of new product features. Should you keep on sampling?

sampling needs time!

you conduct a marketing survey that will guide the development of new product features. Should you keep on sampling?

if you spend too much time sampling, your competitor may develop and start selling their product before you do!

sampling needs time!

you conduct a survey on fashion/ pop music to guide business decisions (e.g. background music for advertisement), Should you keep sampling?

sampling needs time!

you conduct a survey on fashion/ pop music to guide business decisions (e.g. background music for advertisement), Should you keep sampling?

fashion and pop culture evolve quickly (the DGP changes quickly!). If you sample too long, you miss the current trend and act upon a trend that is already outdated!

sampling needs time!

you conduct a student survey in 2019, but few students participate. Now it is April 2020, should you keep sampling?

sampling needs time!

you conduct a student survey in 2019, but few students participate. Now it is April 2020, should you keep sampling?

after March 11 (mid March) covid measures lead to school closures. The DGP has definitively changed!

sampling needs time!

you conduct a survey on young men's opinion about the military in the early 90s. Now it is late 90s. Should you keep sampling?

sampling needs time!

you conduct a survey on young men's opinion about the military in the early 90s. Now it is late 90s. Should you keep sampling?

In the European part of Netherlands, compulsory attendance has been officially suspended since 1 May 1997. Between 1991 and 1996, the Dutch armed forces phased out their conscript personnel and converted to an all-volunteer force. The DGP has definitively changed!

sampling needs time!

you conduct a survey on how to price a drug for which you have a patent. Should you keep on sampling

sampling needs time!

you conduct a survey on how to price a drug for which you have a patent. Should you keep on sampling

even if you do not fear competition, new medical findings may make your drug redundant. You need to sell it asap to make use of your patent!

recap

- we have learned what happens to the sample mean when N gets large: if the sample is representative, it converges to the population mean
- we have also learned that the sample mean is a random variable and approximately normally distributed
- we saw why we like large N: we can shrink the estimator variance
- we carefully looked at situation in which tolerated small N is better than continuing to sample