Confidence Intervals

Data Analysis for Psychology in R 1

Semester 2, Week 1

Dr Umberto Noè

Department of Psychology
The University of Edinburgh

Learning objectives

1. Understand the importance of a confidence interval.
2. Understand the link between standard errors and confidence intervals.
3. Understand how to construct a confidence interval for an unknown parameter of interest.

Part A

Recap

Normal distribution

$X \sim N(\mu, \sigma)$

Probability to the LEFT of a value x:

```
p <- pnorm(x, mean = mu, sd = sigma)
```

Value x having a probability of p to its LEFT:

```
x <- qnorm(p, mean = mu, sd = sigma)
```

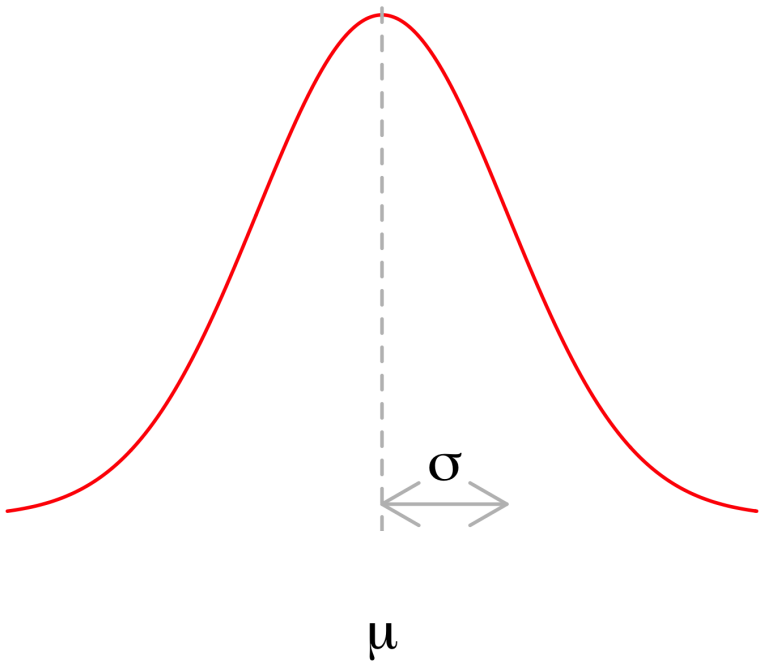
Example with $N(0, 1)$:

```
qnorm(0.975)
```

```
## [1] 1.96
```

```
pnorm(1.96)
```

```
## [1] 0.975
```



Standardisation / z-scoring

- Let $X \sim N(\mu, \sigma)$
- Define:

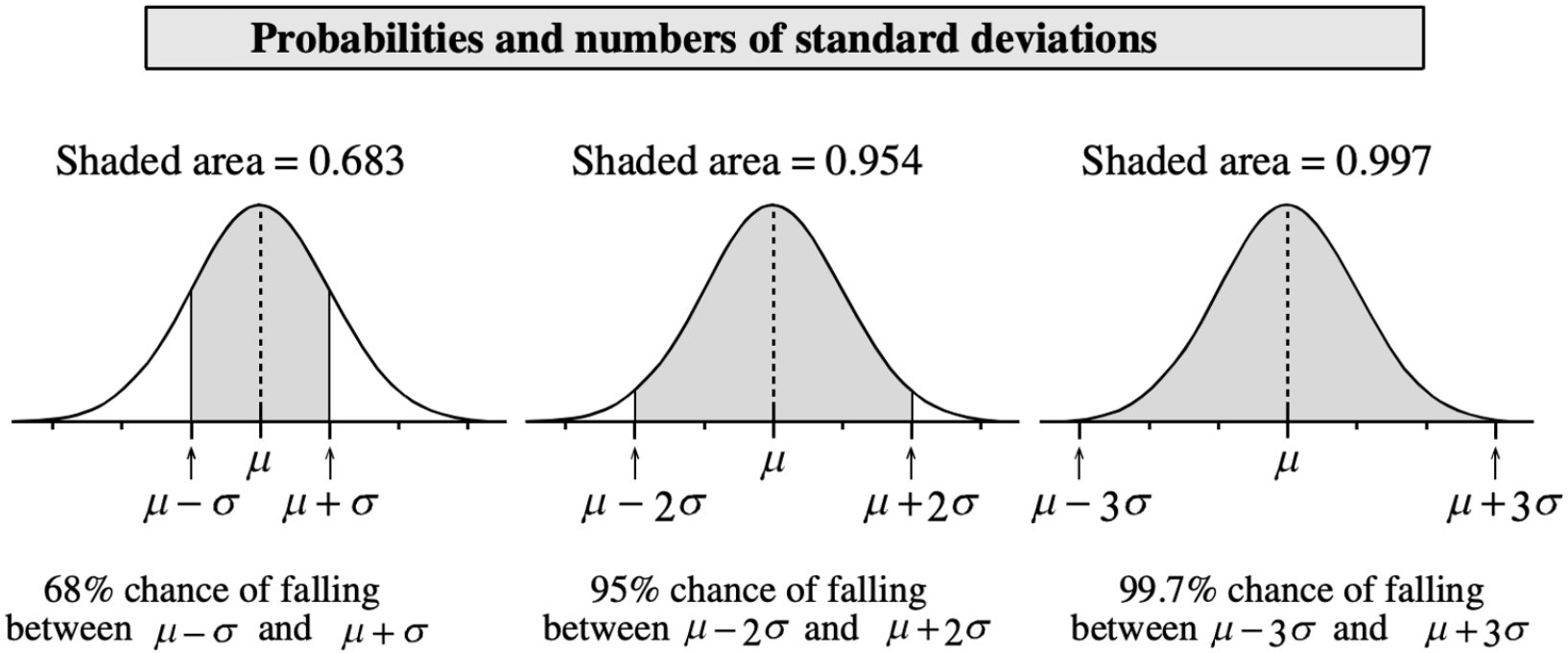
$$Z = \frac{X - \mu}{\sigma}$$

- $Z \sim N(0, 1)$ follows a standard normal distribution
- To transform Z into X we use this transformation:

$$X = \mu + Z \cdot \sigma$$

Normal 68–95–99.7 rule

- Recall that for a random variable $X \sim N(\mu, \sigma)$, roughly 95% of the values fall between $\mu - 2\sigma$ and $\mu + 2\sigma$:



Normal 68–95–99.7 rule

- The interval below contains **roughly** 95% of the values in the distribution:

$$[\mu - 2 \cdot \sigma,\ \mu + 2 \cdot \sigma]$$

- To be more accurate, we need to find the x-values (quantiles) that have 0.025 probability to the left and 0.025 probability to the right, leaving 0.95 probability in the middle.

```
qnorm(c(0.025, 0.975))  # using a N(0,1) distribution
```

```
## [1] -1.96  1.96
```

- The values -1.96 and 1.96 are the quantiles of a standard Normal distribution, cutting a probability of 0.025 in the tails each.

$$\begin{aligned} z = -1.96 \quad &\rightarrow \quad x = \mu - 1.96 \cdot \sigma \\ z = 1.96 \quad &\rightarrow \quad x = \mu + 1.96 \cdot \sigma \end{aligned}$$

- The correct interval comprising exactly 95% of the values is:

$$[\mu - 1.96 \cdot \sigma,\ \mu + 1.96 \cdot \sigma]$$

Using statistics to estimate parameters

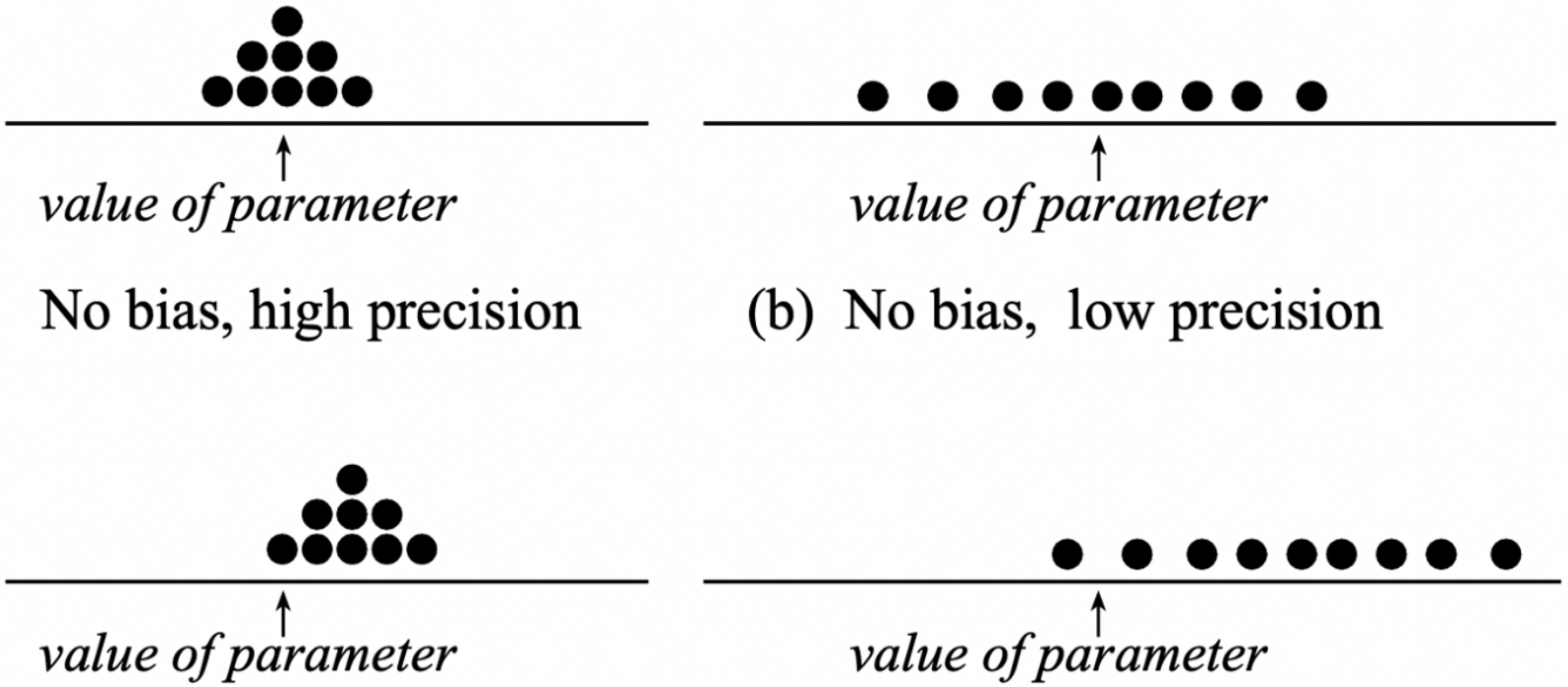
- Without loss of generality, in this lecture we will focus on the mean as the numerical summary of data.
  - Population mean $\rightarrow$ unknown $\rightarrow$ example of a parameter $\rightarrow \mu$
  - Sample mean $\rightarrow$ we can compute it $\rightarrow$ example of a statistic $\rightarrow \bar{x}$
- We are typically interested in **estimating an unknown population mean** (a **parameter**, μ) using the **mean computed on a random sample** (a **statistic**, $\bar{x}$).
- We will also call the sample mean (= statistic) the **estimate**.
- FACT: statistics vary from sample to sample and have a **sampling distribution**.
- The standard deviation of the sampling distribution is called the **standard error** (SE)
  - SE tells us the size of the typical "estimation error" = $\bar{x} - \mu$.
  - $SE = \frac{\sigma}{\sqrt{n}}$

Key question

> How accurate is our estimate?

- We are interested in how accurate our statistic $\bar{x}$ is as an estimate of the unknown parameter μ.
- Accuracy is a combination of two things:
  - No bias
  - Precision
- We avoid bias if we use random sampling. We have bias if our samples systematically do not include a part of the population.
  - If you choose convenience samples, you will systematically over-estimate or under-estimate the true value.
- Precision relates to the variability of the sampling distribution, and the Standard Error (SE) is used to quantify precision.
  - SE = sd(of sampling distribution)
  - The smaller the SE, the closer our sample mean will tend to be to the population mean

Bias vs Precision



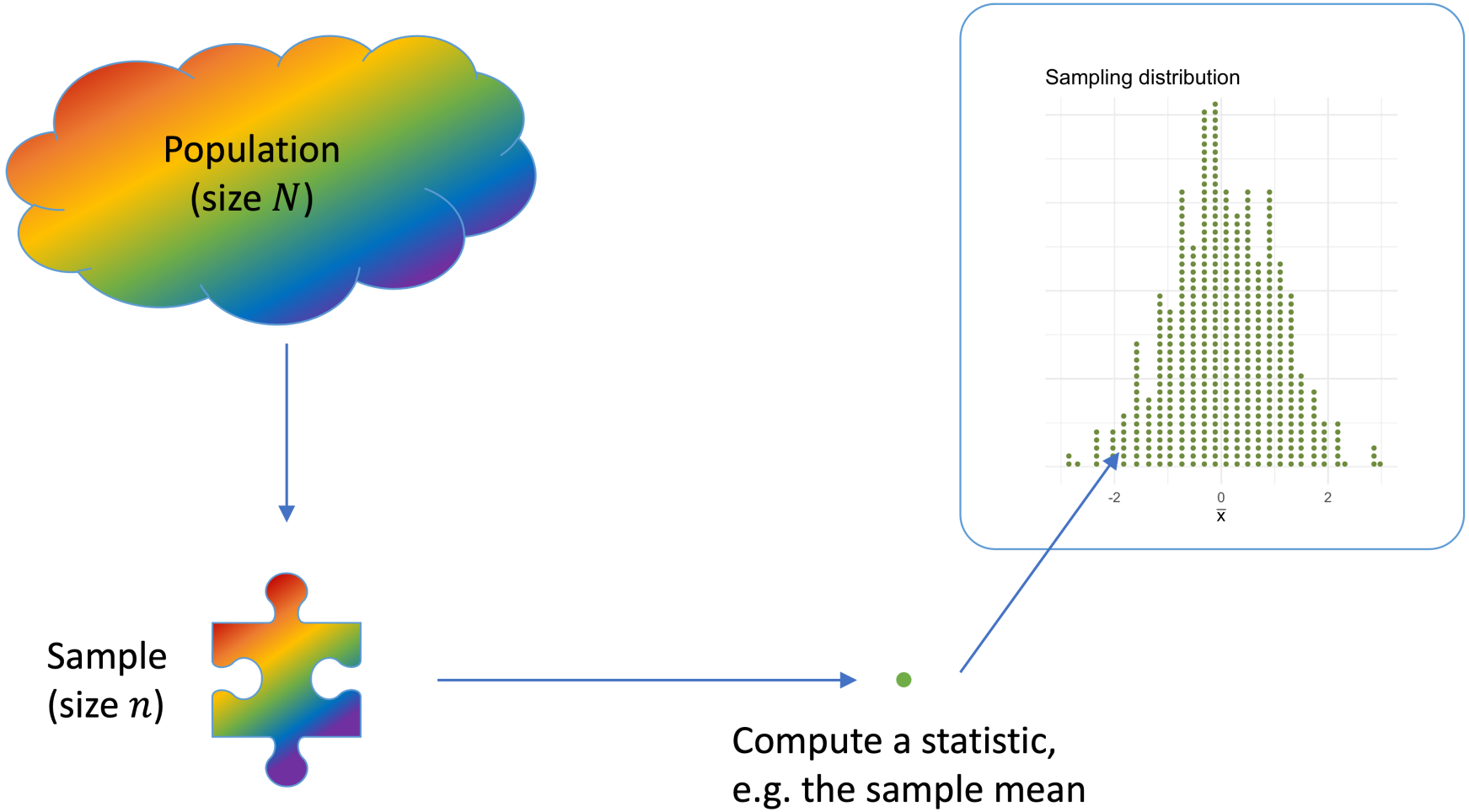
(a) No bias, high precision

(b) No bias, low precision

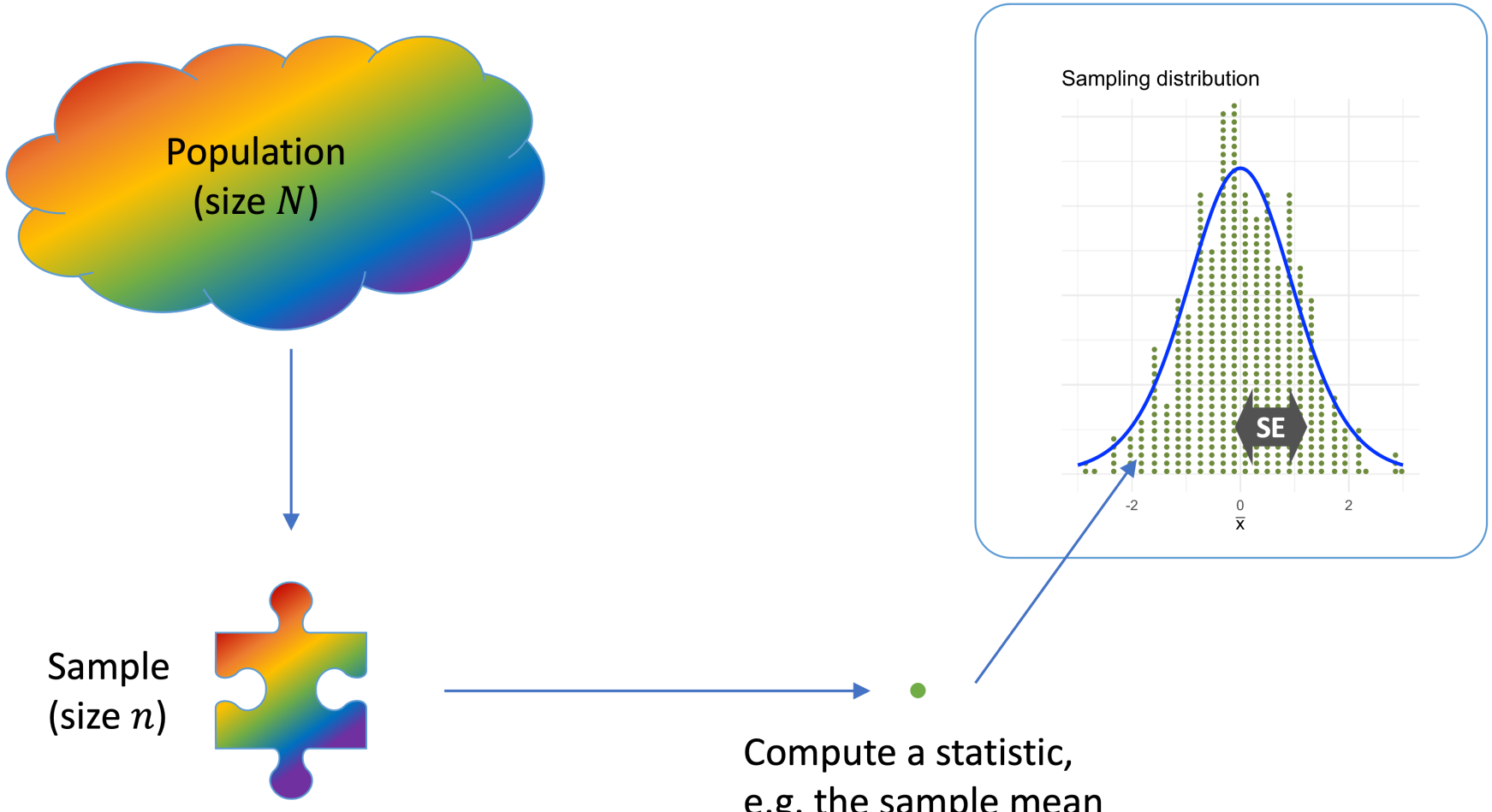
(c) Biased, high precision

(d) Biased, low precision

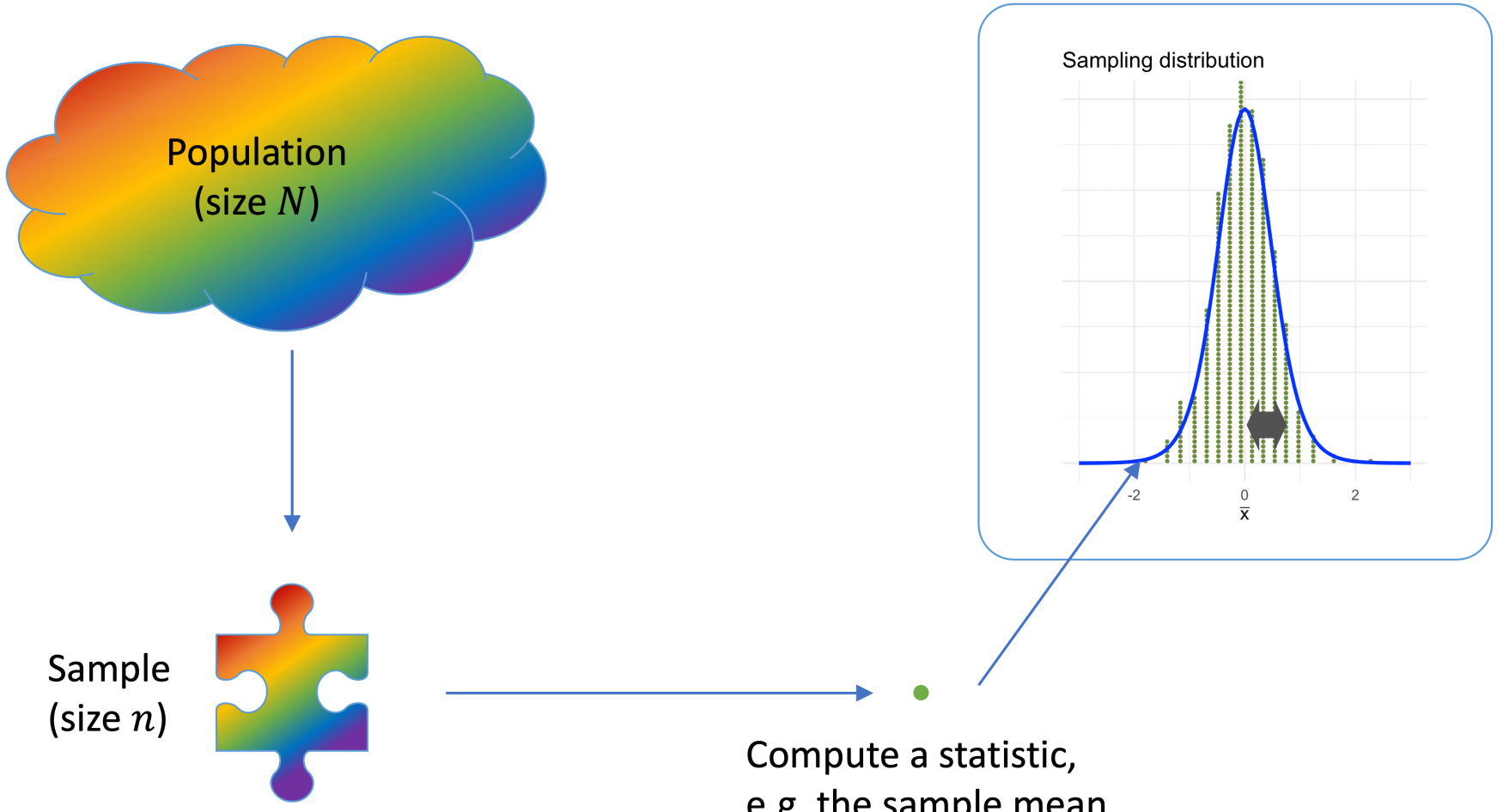
Sampling distribution



Sampling distribution



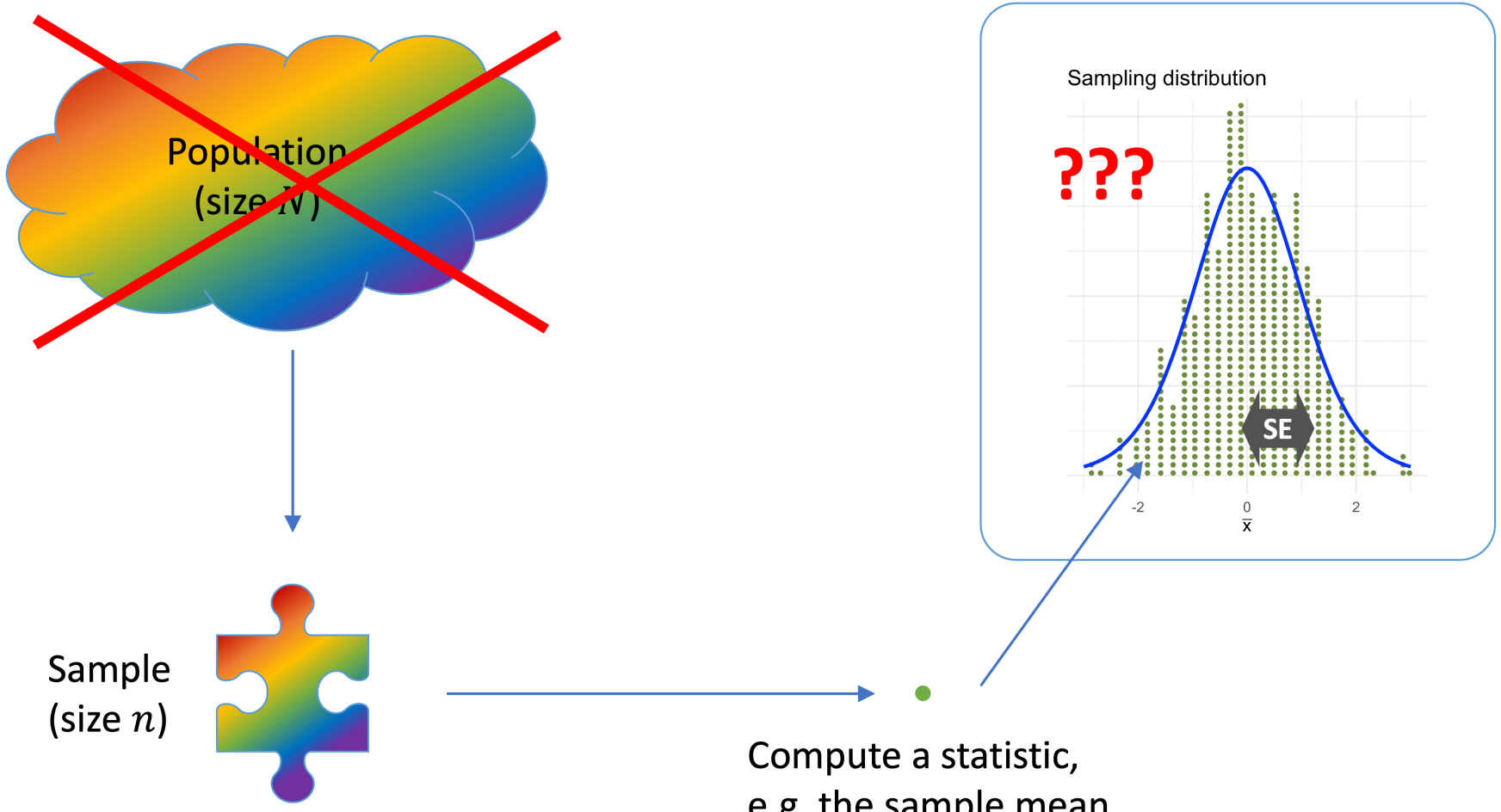
Sampling distribution



Part B

One sample only

One sample only



One sample only: Precision of sample mean

- If we do NOT have the population data:
  - we cannot compute μ, the population mean
  - we also cannot compute σ, the population standard deviation
- Recall that σ is required to assess the precision of the sample mean by computing the SE:

$$SE = \frac{\sigma}{\sqrt{n}}$$

- How can we compute the SE of the mean if we **do not have data on the full population**, and we **can only afford one sample** of size n?

One sample only: Precision of sample mean

- We must also estimate σ with the corresponding sample statistic.
- Substitute σ with the standard deviation computed in the sample, s.
- Standard error of the mean becomes:

$$SE = \frac{s}{\sqrt{n}}$$

- Report estimate (sample mean), along with a measure of its precision (the above SE).

Part C

Confidence Intervals

Key idea

- Parameter estimate = single number. Almost surely the true value will be different from our estimate.
- Range of plausible values for the parameter, called **confidence interval**. More likely that the true value will be captured by a range.

Point estimate	Confidence interval
	

Confidence interval

- Confidence interval (CI) = range of plausible values for the parameter.
- To create a confidence interval we must decide on a confidence level.
- Confidence level = a number between 0 and 1 specified by us. How confident do you want to be that the confidence interval will contain the true parameter value?
- The larger the confidence level, the wider the confidence interval.
    - How confident are you that I am between 39 and 42 years old?
    - How confident are you that I am between 35 and 50 years old?
    - How confident are you that I am between 18 and 70 years old?
- Typical confidence levels are 90%, 95%, and 99%.

CI for the population mean

- Recall that if $X \sim N(\mu, \sigma)$, 95% of the values are between

$$[\mu - 1.96 \cdot \sigma,\ \mu + 1.96 \cdot \sigma]$$

- The sample mean follows a normal distribution:

$$\bar{X} \sim N(\mu_{\overline{X}}, \sigma_{\overline{X}})$$

where

- $\mu_{\overline{X}} = \mu$
- $\sigma_{\overline{X}} = SE = \dfrac{\sigma}{\sqrt{n}}$
- Substitute in the interval above:

$$\left[\mu - 1.96 \cdot \frac{\sigma}{\sqrt{n}},\ \mu + 1.96 \cdot \frac{\sigma}{\sqrt{n}}\right]$$

Estimates of μ and σ

- Recall that we do not have the full population data. **We can only afford one sample!**
- We don't have the population mean μ and we estimated it with the sample mean $\bar{x}$
- However, we also don't have σ so we need to estimate it with s, the sample standard deviation:

$$\left[\bar{x} - 1.96 \cdot \frac{s}{\sqrt{n}},\ \bar{x} + 1.96 \cdot \frac{s}{\sqrt{n}}\right]$$

- However, this interval is now **wrong!**
- Because we didn't know σ and we had to estimate it with s, this bring and **extra element of uncertainty**
- As we are unsure about the actual value of the population standard deviation, the reference distribution is no longer Normal, but a distribution that is more "uncertain" and places higher probability in the tails of the distribution.
- When the population standard deviation is unknown, the sample mean follows a t-distribution.
- The quantiles -1.96 and 1.96 refer to the normal distribution, so these are wrong and we need to find the correct ones!

t-distribution

t-distribution

- A distribution similar to the standard Normal distribution, also with a zero mean
- Depends on a number called **degrees of freedom** (DF) = sample size - 1. That is, $df = n - 1$.
- We write the distribution as:

$$t(n-1)$$

- Suppose the sample size is 20. In R:

```
qt(0.025, df = 19)    # quantile = t-value with 0.025 prob to the LEFT
```

```
## [1] -2.093
```

```
pt(-2.093, df = 19)   # prob to the LEFT of t = -2.093
```

```
## [1] 0.025
```

Finally: the correct confidence interval

- Now we can finally compute the correct confidence interval.
- We need to replace the quantiles with those from the $t(n-1)$ distribution, denote them by $-t^*$ and $+t^*$, and these will be different all the time as they depend on the sample size.
- Generic form the of the CI for the mean:

$$\left[\bar{x} - t^* \cdot \frac{s}{\sqrt{n}},\ \bar{x} + t^* \cdot \frac{s}{\sqrt{n}}\right]$$

- Generic form the of the 95% CI for the mean with a sample of size $n = 20$:

```
qt(c(0.025, 0.975), df = 20 - 1)
```

```
## [1] -2.093  2.093
```

$$\left[\bar{x} - 2.093 \cdot \frac{s}{\sqrt{n}},\ \bar{x} + 2.093 \cdot \frac{s}{\sqrt{n}}\right]$$

Other confidence levels

- Generic form the of the 99% CI for the mean with a sample of size $n = 20$:

```
qt(c(0.005, 0.995), df = 20 - 1)
```

```
## [1] -2.861  2.861
```

$$\left[\bar{x} - 2.861 \cdot \frac{s}{\sqrt{n}},\ \bar{x} + 2.861 \cdot \frac{s}{\sqrt{n}}\right]$$

Example: 95% CI for the pop. mean salary

- Parameter of interest: mean yearly salary of a NFL player in the year 2019, denoted μ.
- Sample of 50 players:

```
library(tidyverse)
nfl_sample <- read_csv("https://uoepsy.github.io/data/NFLSample2019.csv")
dim(nfl_sample)
```

```
## [1] 50  5
```

```
head(nfl_sample)
```

```
## # A tibble: 6 × 5
##   Player          Position Team    TotalMoney YearlySalary
##   <chr>           <chr>    <chr>        <dbl>        <dbl>
## 1 Najee Goode     43OLB    Jaguars      0.805        0.805
## 2 Jack Crawford   43DT     Falcons      9.9          2.48
## 3 Tra Carson      RB       Lions        1.23         0.615
## 4 Jordan Richards S        Ravens       0.805        0.805
## 5 Desmond Trufant CB       Falcons     68.8         13.8
## 6 Alex Anzalone   43OLB    Saints       3.47         0.866
```

Example: 95% CI for the pop. mean salary

```
xbar <- mean(nfl_sample$YearlySalary)
xbar
```

```
## [1] 3.359
```

```
s <- sd(nfl_sample$YearlySalary)
s
```

```
## [1] 4.312
```

```
n <- nrow(nfl_sample)
n
```

```
## [1] 50
```

```
SE <- s / sqrt(n)
SE
```

```
## [1] 0.6098
```

```
qt(c(0.025, 0.975), df = n-1)
```

```
## [1] -2.01  2.01
```

```
xbar - 2.01 * SE
```

```
## [1] 2.133
```

```
xbar + 2.01 * SE
```

```
## [1] 4.584
```

Example: 95% CI for the pop. mean salary

- The 95% confidence interval for the mean salary of **all** NFL players in the year 2019 is [2.13, 4.58] million dollars.
- Write this up as:

> We are 95% confident that the average salary of a NFL player in 2019 is between 2.13 and 4.58 million dollars.

Part D

Warning on interpretation

Warning!

- If you had many random samples and computed a 95% confidence interval from each sample:
  - about 95% of those intervals will contain the true parameter value
  - about 5% of those intervals will **not** contain the true parameter value
- Example 1: if you had 100 random samples and computed a 95% confidence interval from each sample:
  - about 95 (= 100 * 0.95) of those intervals will contain the true parameter value
  - about 5 (= 100 * 0.05) of those intervals will **not** contain the true parameter value
- Example 2: if you had 20 random samples and computed a 95% confidence interval from each sample:
  - about 19 (= 20 * 0.95) of those intervals will contain the true parameter value
  - about 1 (= 20 * 0.05) of those intervals will **not** contain the true parameter value

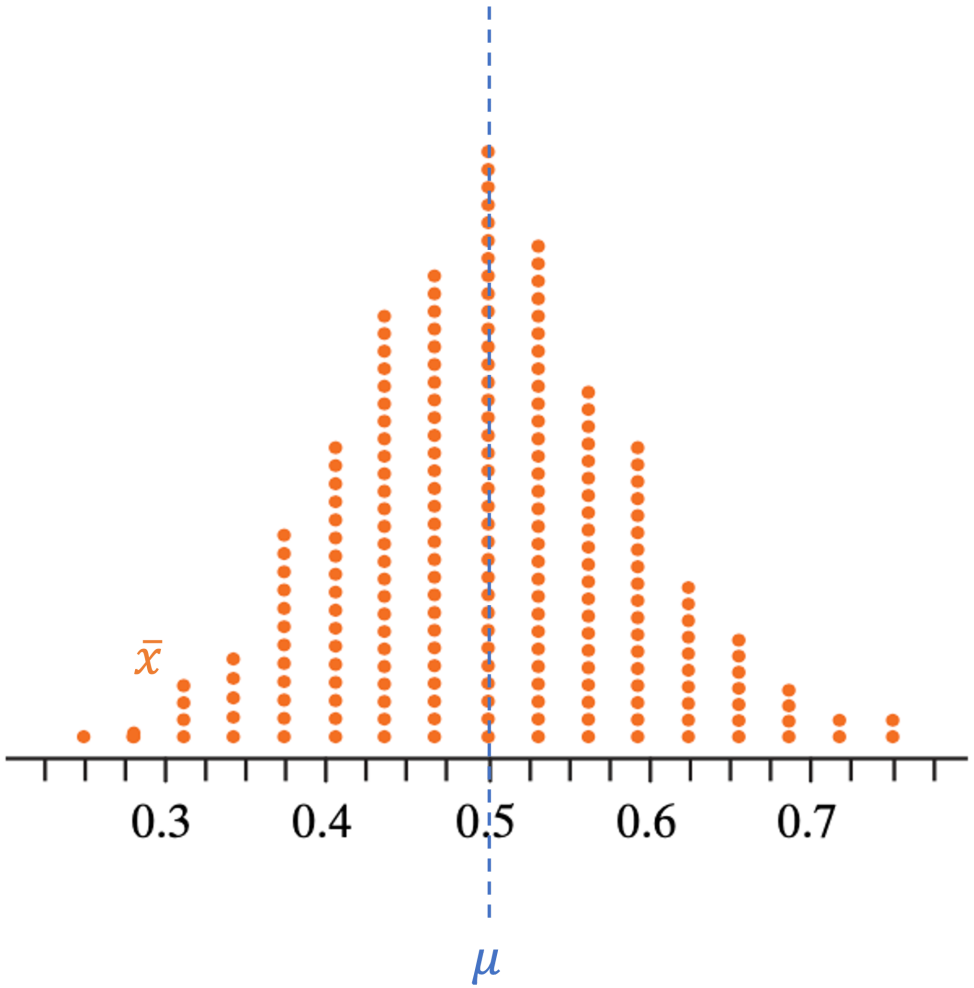
Warning!

- Consider again example 2, where you have 20 random samples and built a confidence interval from each sample.
- We speak about **probability** when we refer to the **collection** of those 20 confidence intervals.
  That is, the probability the that **collection** of confidence intervals will contain the true parameter value is 0.95.
    - Think of this as

$$\frac{\text{number of CIs containing } \mu}{\text{total number of CIs}} = \frac{19}{20} = 0.95$$

- We speak of **confidence** when we refer to just **one** confidence interval that we have computed.
  Say the 95% CI is [2.5, 5.3] min. We would say: we are 95% confident that the population mean is between 2.5 and 5.3 minutes.
    - It is **wrong** to say that there is a 95% probability that the population mean is between 2.5 and 5.3 minutes.

Warning!



Warning!