

Classical statistical inference

Part 2

Associated notebook:

[04-Basic statistical inference frequentists 2/Frequentist inference 01.ipynb](#)

What is inference?

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)



Derive **INFORMATION** based on **DATA**

Examples:

- Exoplanet transit $\Rightarrow M, d \Rightarrow P(M | d)$
- Supernovae distances \Rightarrow Expansion rate H_0

Inference generally implies an underlying *statistical model*: PDF or regression laws with *parameters* θ

FREQUENTIST STATISTICIAN:



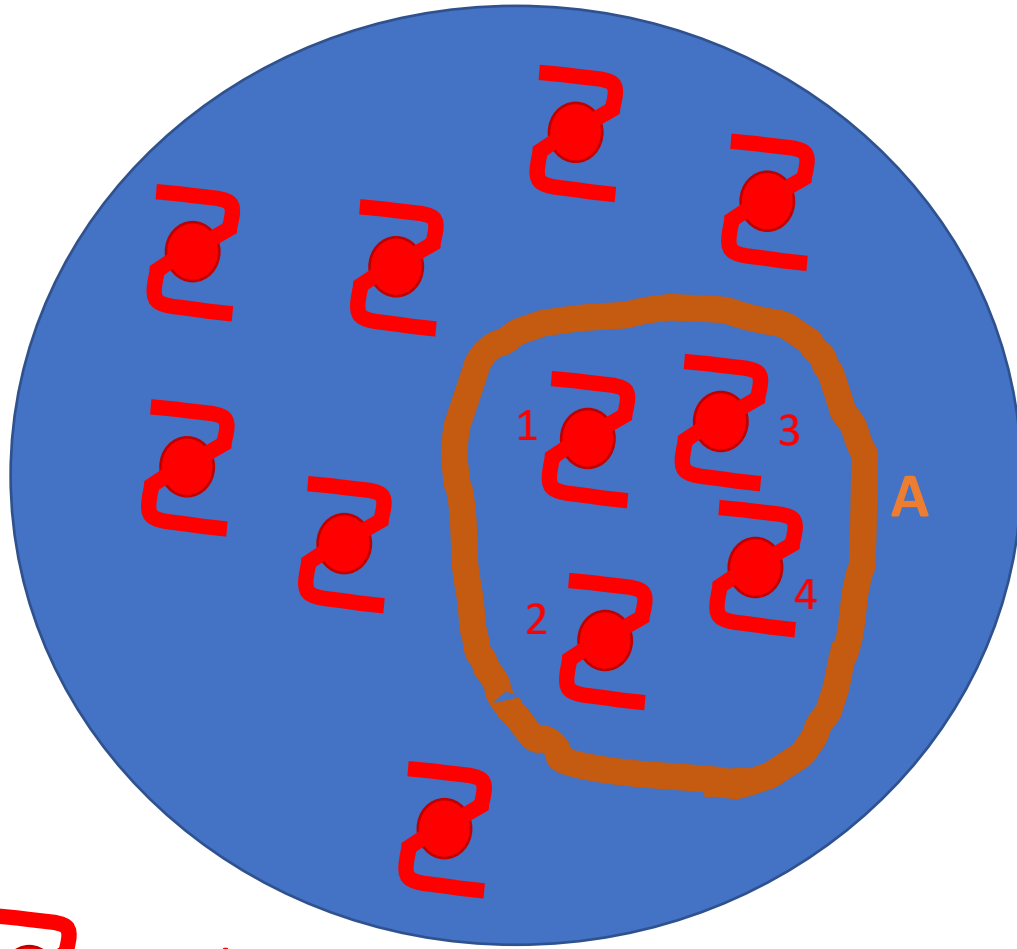
BAYESIAN STATISTICIAN:



Three types of inference:

- Point estimation: “best” θ
- Confidence interval: Confidence around θ
- Hypothesis testing: *data OK w. model?*

Point estimate $\hat{\theta}$



= galaxy

Example:

θ = mean mag. of a population of galaxies

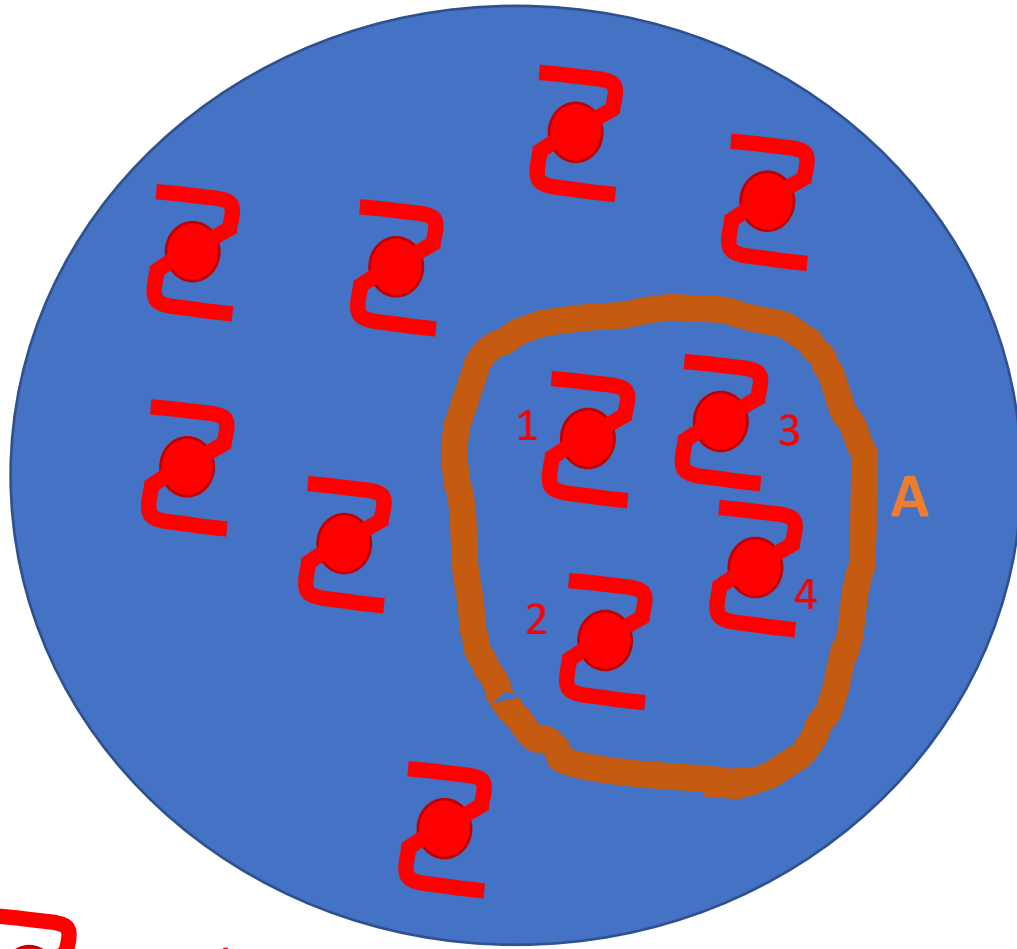
A: Your data set = subsample of measurements:

$A = \{X_1, X_2, X_3, X_4\}$ where X = mag. (this is a RV)

$$\hat{\theta} = \frac{X_1 + X_2 + X_3 + X_4}{4} \equiv \text{Point estimate of } \theta$$

If you do the experiment with another sample
(\Rightarrow different *realisation*) you will get another $\hat{\theta}$

Point estimate $\hat{\theta}$



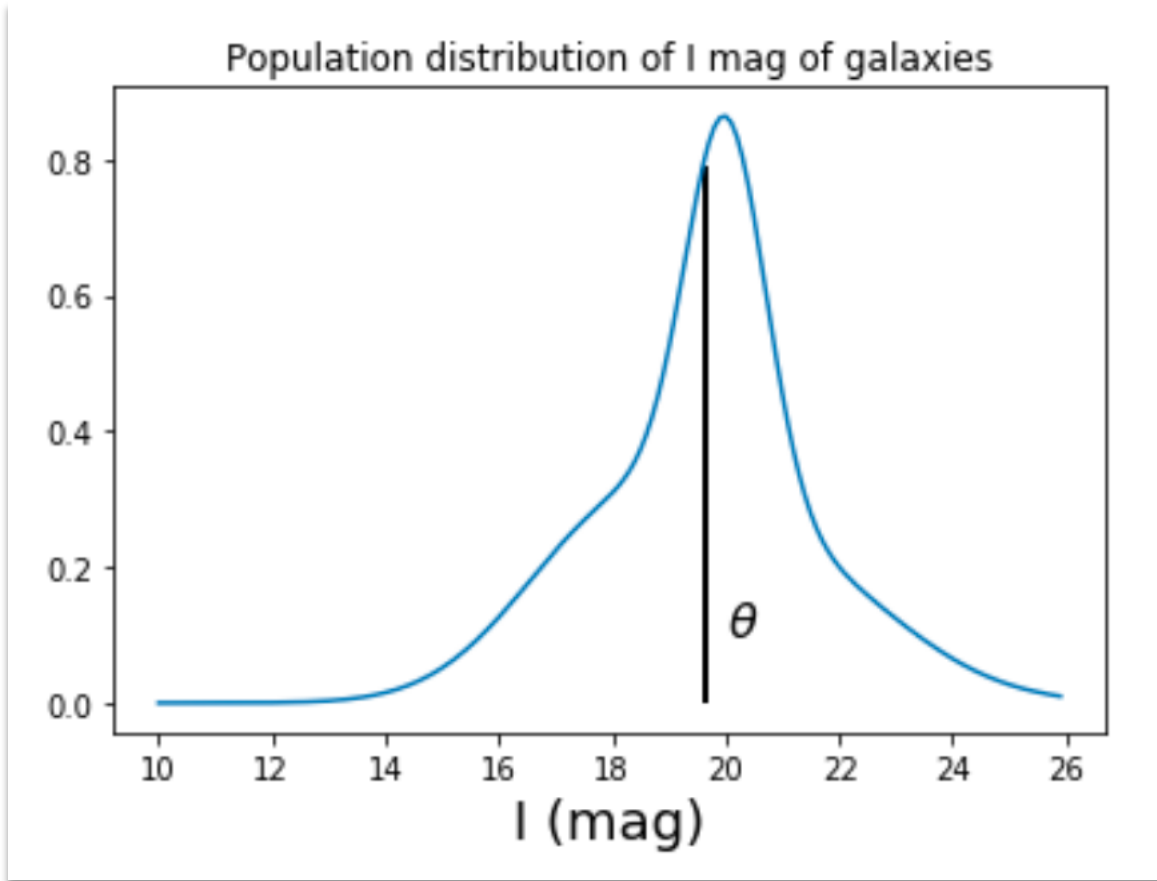
Generalisation:

$$\hat{\theta} = g(X_1, X_2, X_3, \dots, X_n)$$

- Point estimate of a param. is a *function* of RV X_1, \dots
- It is as well a **Random Variable (RV)**
- It can be biased, is characterized by a variance but should ideally be consistent (converges towards θ)
- Distribution of $\hat{\theta}$ is called *sampling distribution*



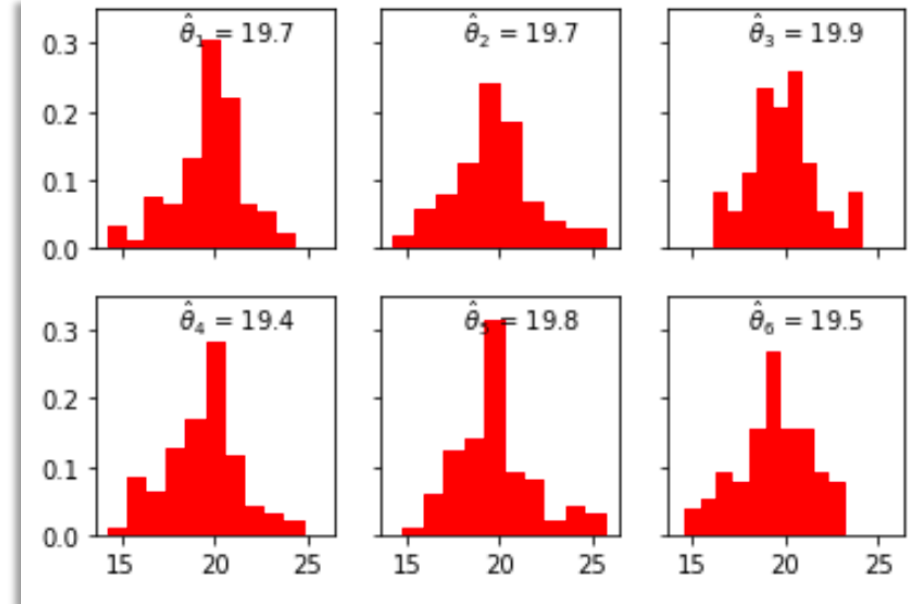
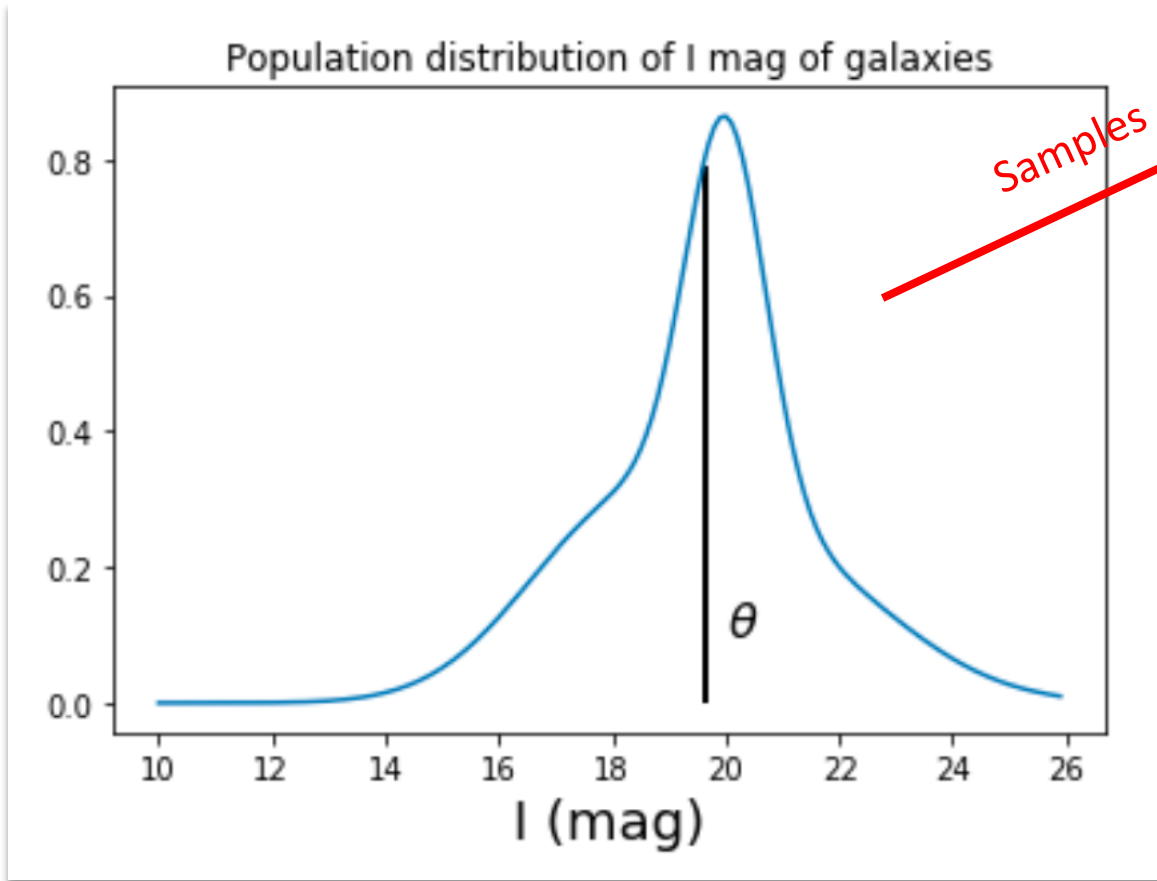
Point estimate $\hat{\theta}$



Population mean: $\theta = 19.66$

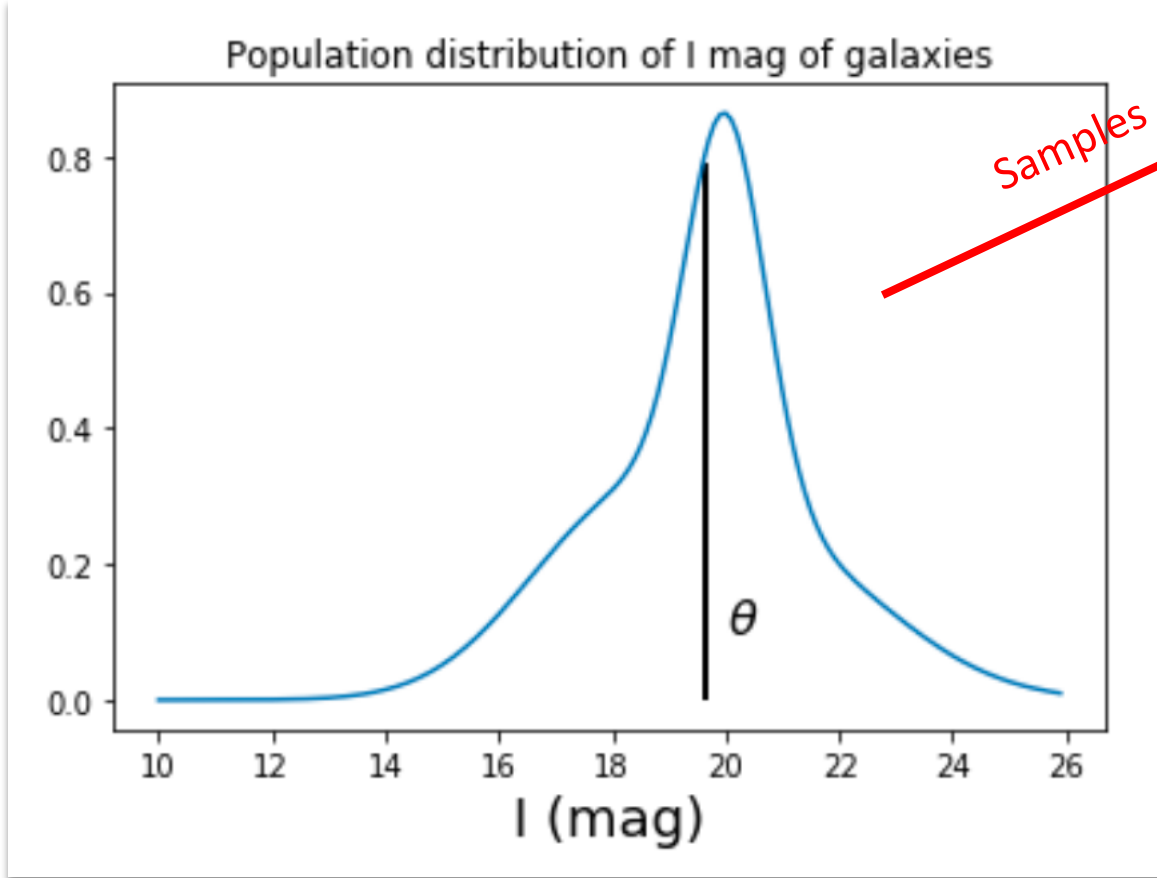
Six different *samples* drawn from the true population

Point estimate $\hat{\theta}$



Population mean: $\theta = 19.66$

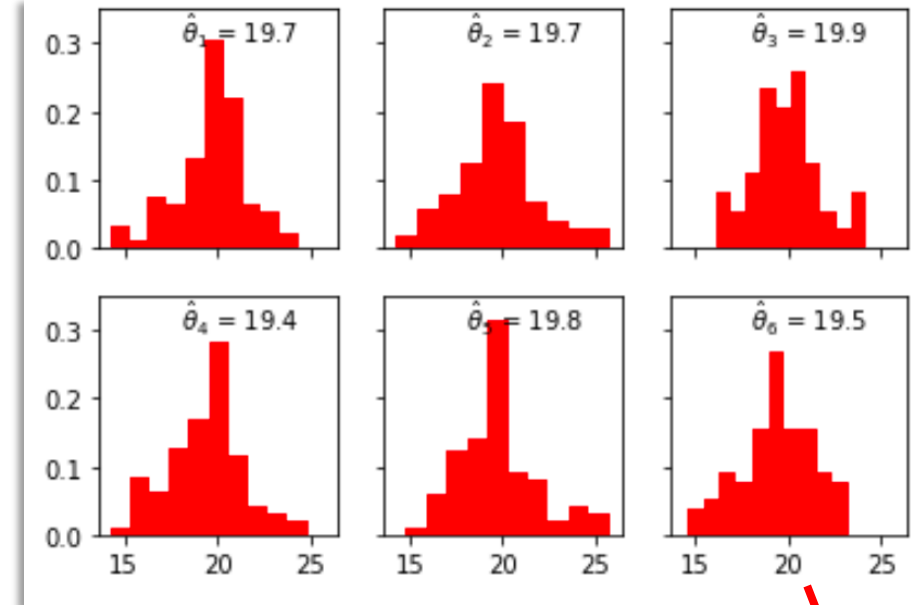
Point estimate $\hat{\theta}$



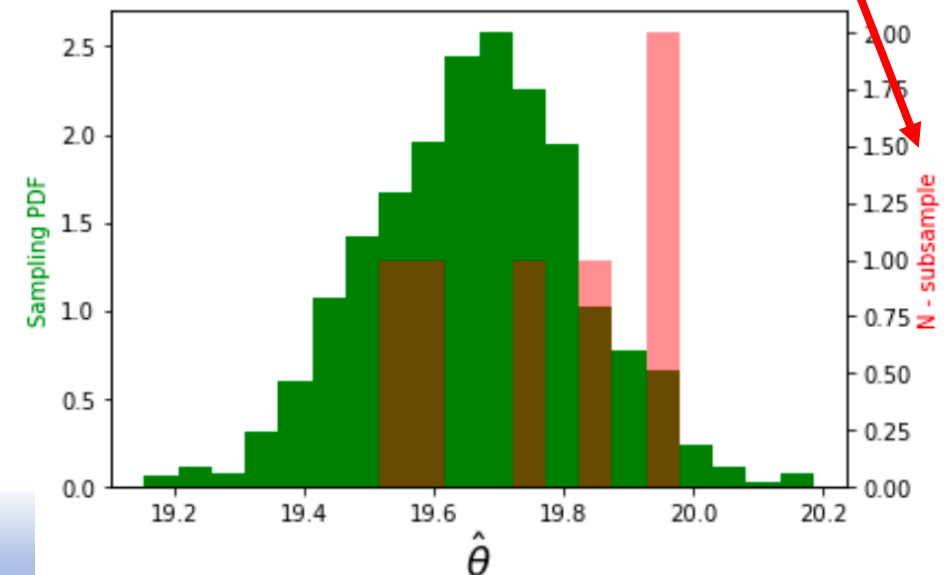
Population mean: $\theta = 19.66$

Go to: Sect. II.1 of the notebook

Six different *samples* drawn from the true population



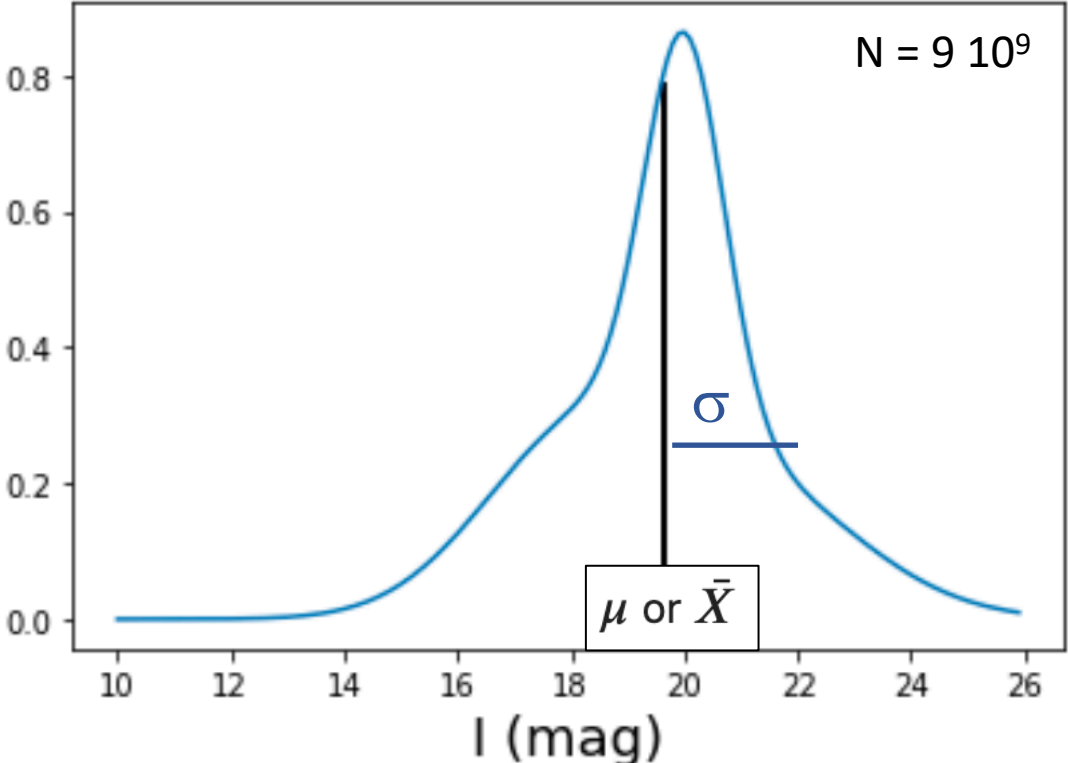
Distribution of $\{\hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_k\} =$ sample distribution



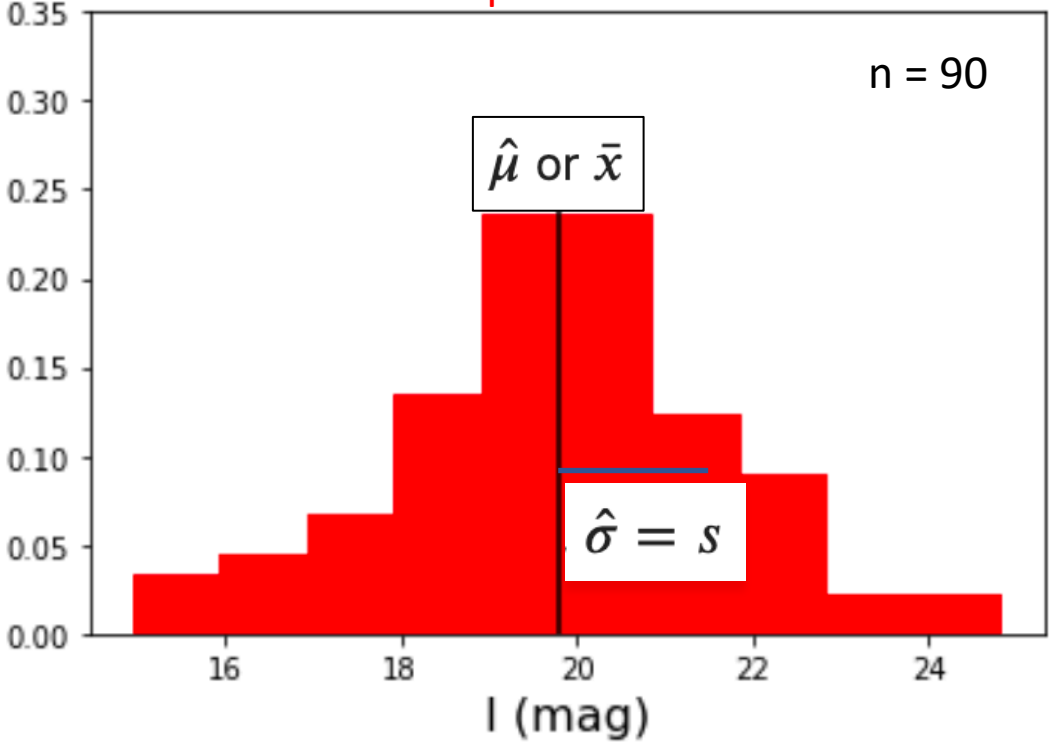
Summary statistics

Name	Population Statistics	Sample Statistics
size	N	n
mean	$\mu = \bar{X} = \frac{\sum_i X_i}{N}$	$\hat{\mu} = \bar{x} = \frac{\sum_i x_i}{n}$
Variance	$\sigma^2 = \frac{\sum_i (X_i - \bar{X})^2}{N}$	$s^2 = \frac{\sum_i (x_i - \bar{x})^2}{n-1}$
Standard deviation	$\sigma = \sqrt{\sigma^2}$	$\hat{\sigma} = s = \sqrt{s^2}$

Population distribution of I mag of galaxies



Sample statistics



Summary (sample) statistics: *standard error*

Standard error (stde) \neq Standard deviation (std)

Name	Formula
Standard error on the mean	$stde(\bar{x}) = \frac{s}{\sqrt{n}}$
Standard error on the stdev	$stde(s) = s/\sqrt{2(n-1)}$
Standard error on proportions	$stde(\hat{p}) = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$

Central limit theorem


When **independent random variables** are added, their *sum* tends towards a normal distribution (if $n \gg$)


- This is true even if the original RV are not normally distributed
- Sampling dist. of mean tends (for large n) towards a Normal distribution
- Sampling dist. of variance (for large n) does NOT tend towards a Normal distribution

Go to: Sect. II.1.1. of the notebook

Distribution of estimators

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$

- Sample distribution of $\hat{\mu} \sim N(\mu, \sigma/\sqrt{n}) \iff Z = \frac{\bar{X} - \hat{\mu}}{(\sigma/\sqrt{n})} \sim N(0, 1)$
- Sample distribution of $t = \frac{\bar{X} - \hat{\mu}}{(s/\sqrt{n})} \sim t(n-1)$ s is derived from the sample dist.


Student distribution
- Sample distribution of $S = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$.


Chi square distribution

Classical statistical inference: *confidence interval*

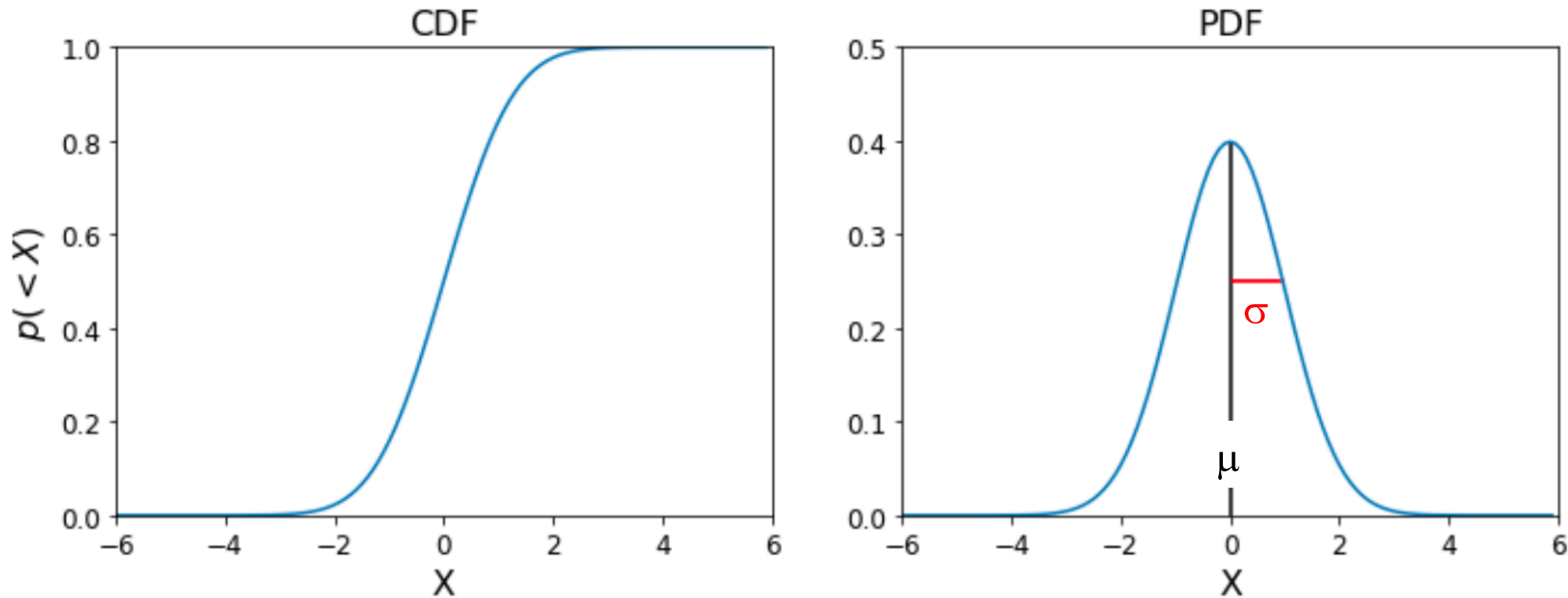
Part 3

Associated notebook:

[04-Basic statistical inference frequentists 2/Frequentist inference 02.ipynb](#)

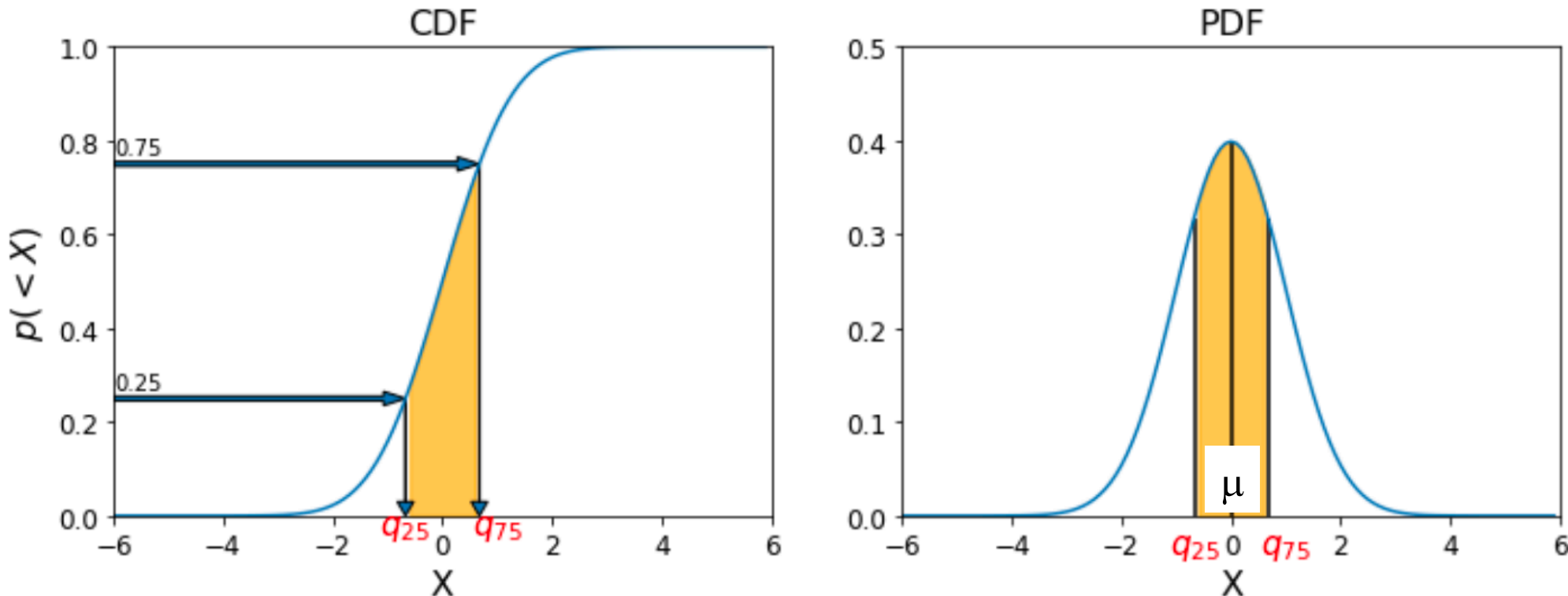
Confidence interval on a RV (descriptive stat)

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



Confidence interval on a RV (descriptive stat)

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$

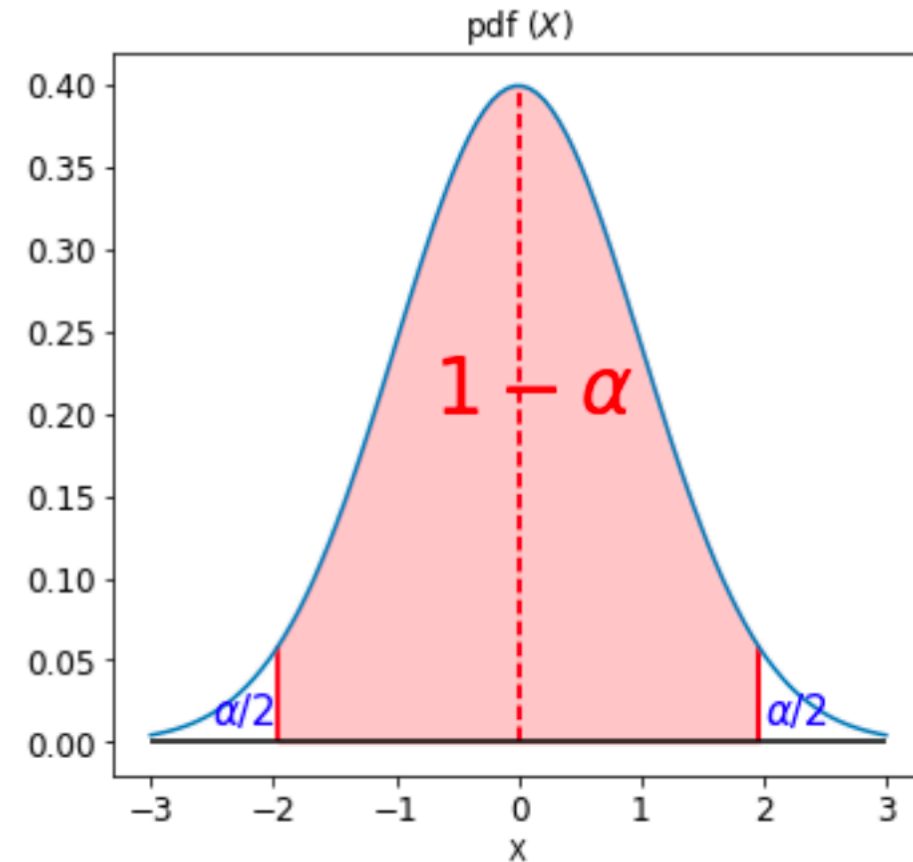
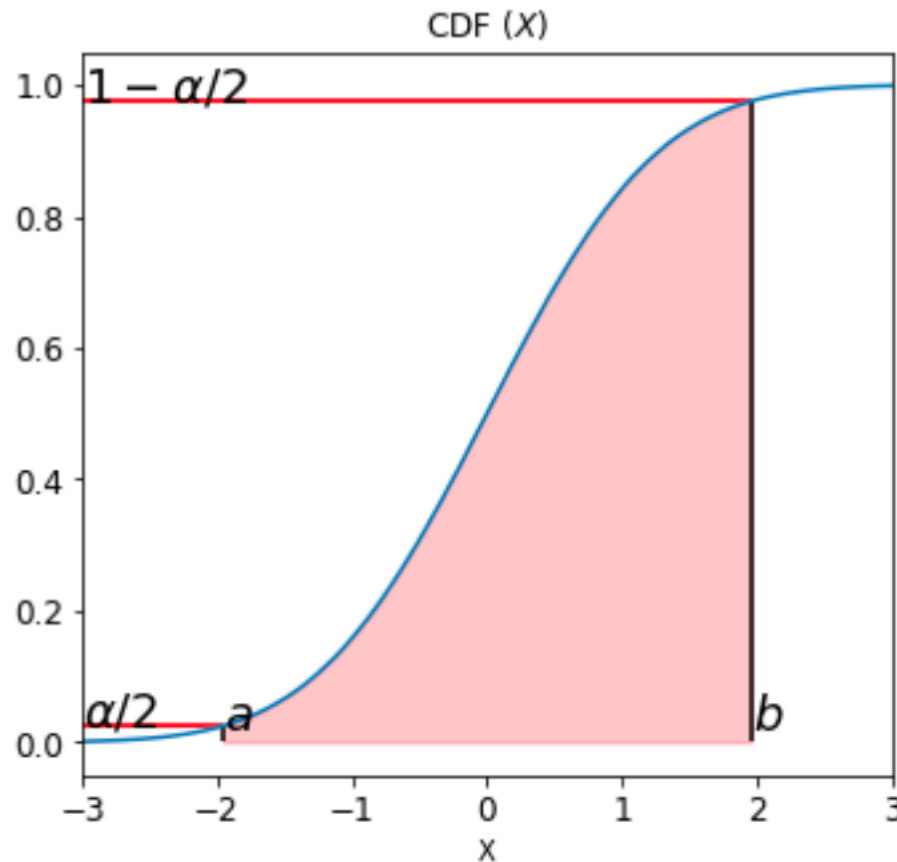


Interquartile IQ = $[q_{25}, q_{75}] \equiv 50\% \text{ CI around } \mu$

$$1 - \alpha \text{ CI} \equiv [a, b] \text{ such that } p(a \leq X \leq b) = (1 - \alpha)$$

Confidence interval on a RV (descriptive stat)

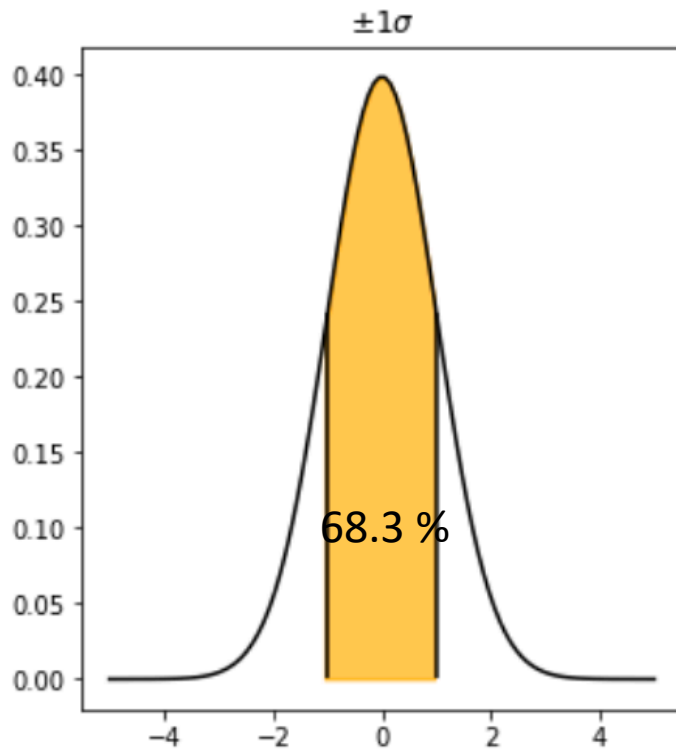
If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



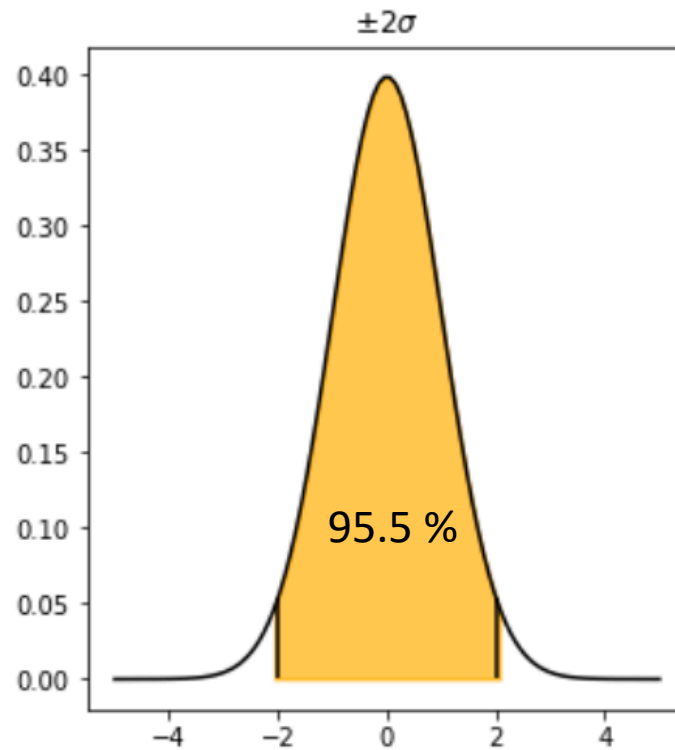
$$1 - \alpha \text{ CI} \equiv [a, b] \text{ such that } p(a \leq X \leq b) = (1 - \alpha)$$

Confidence interval on a RV (descriptive stat)

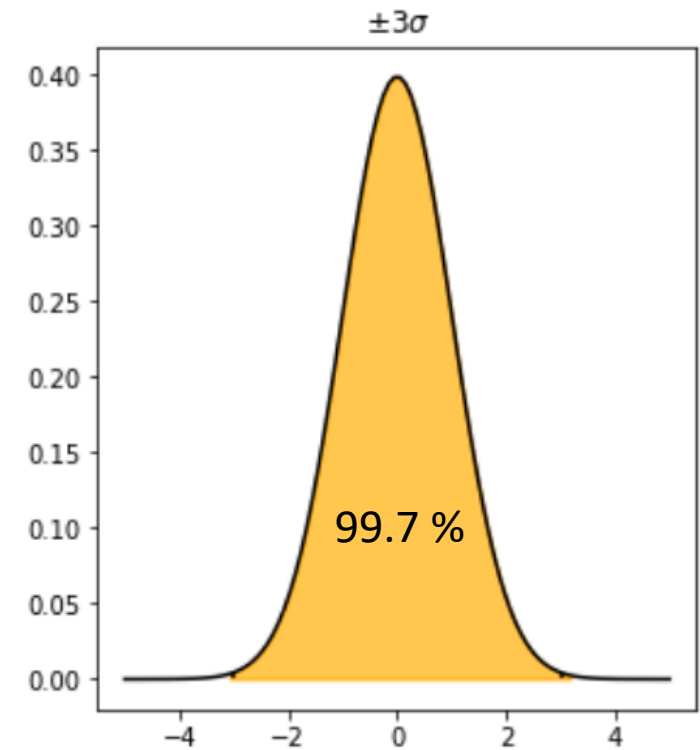
If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



$$[\mu - 1\sigma, \mu + 1\sigma] \equiv 68.3\%$$



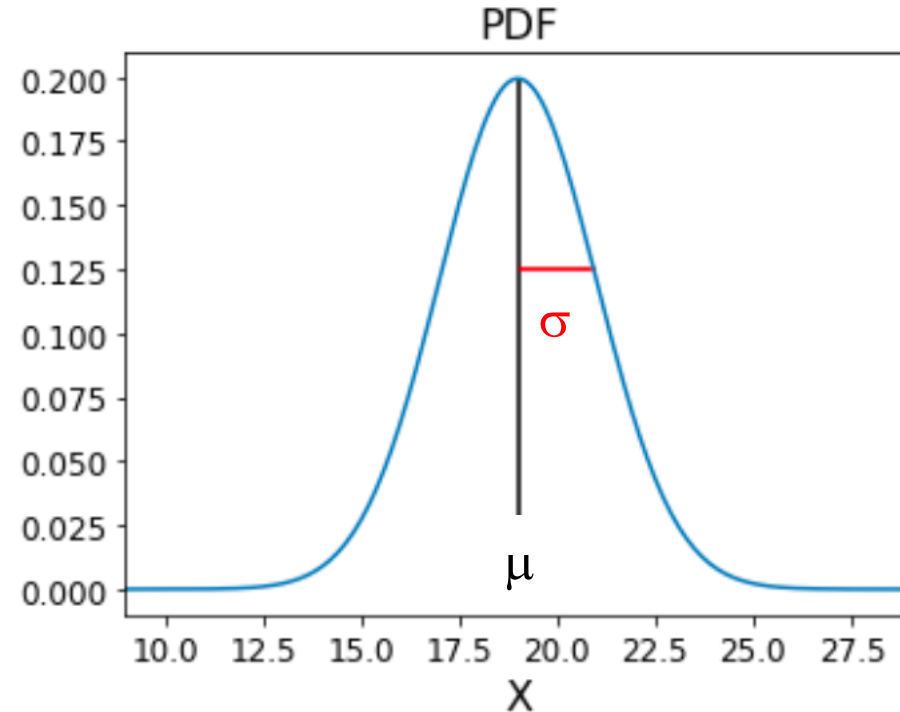
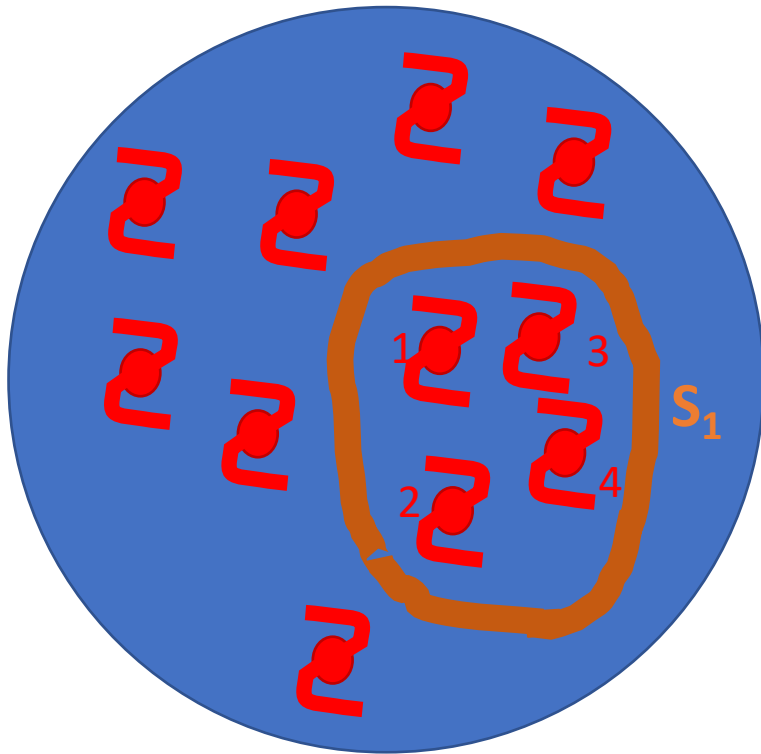
$$[\mu - 2\sigma, \mu + 2\sigma] \equiv 95.5\%$$



$$[\mu - 3\sigma, \mu + 3\sigma] \equiv 99.7\%$$

Confidence interval on $\hat{\mu}$

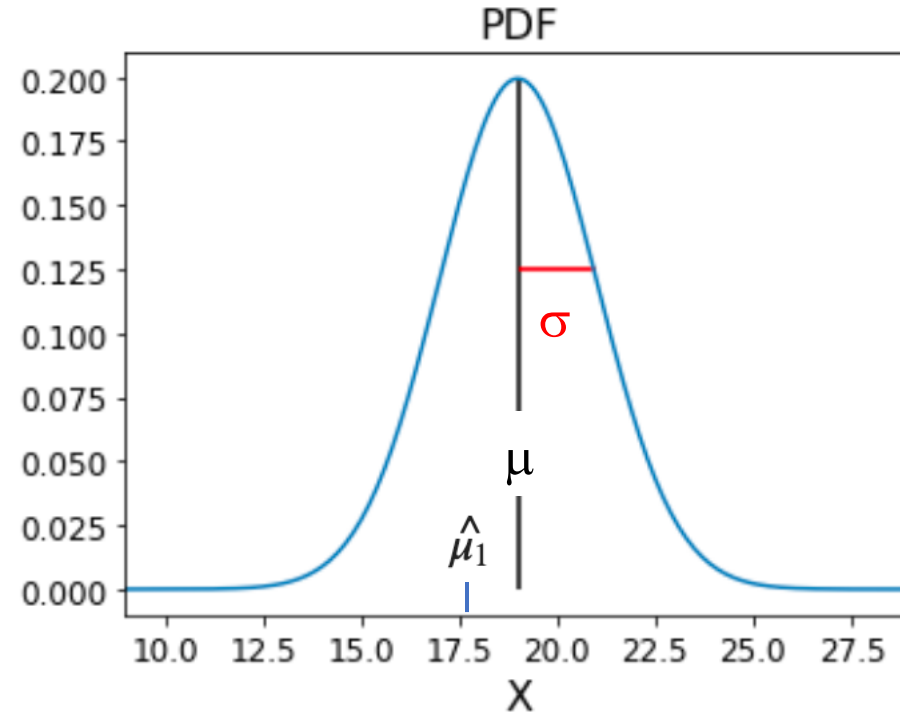
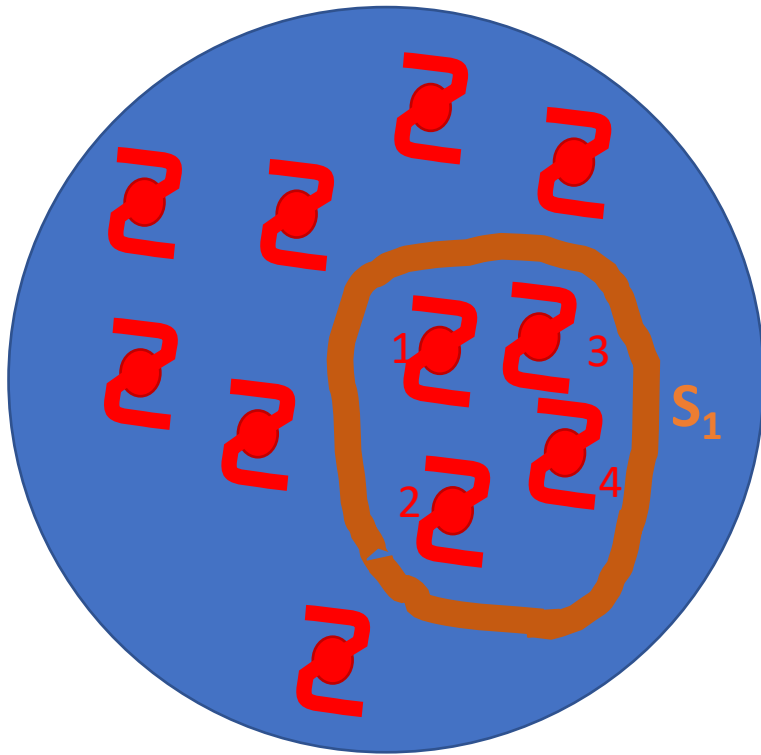
If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



$$S_1 = \{x_1, x_2, x_3, x_4\}$$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



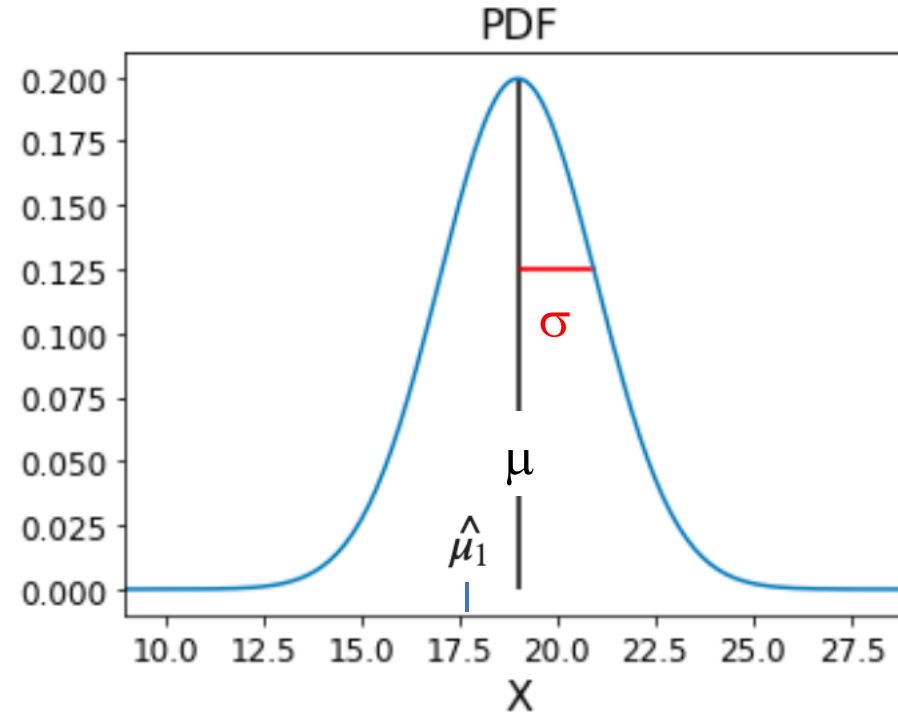
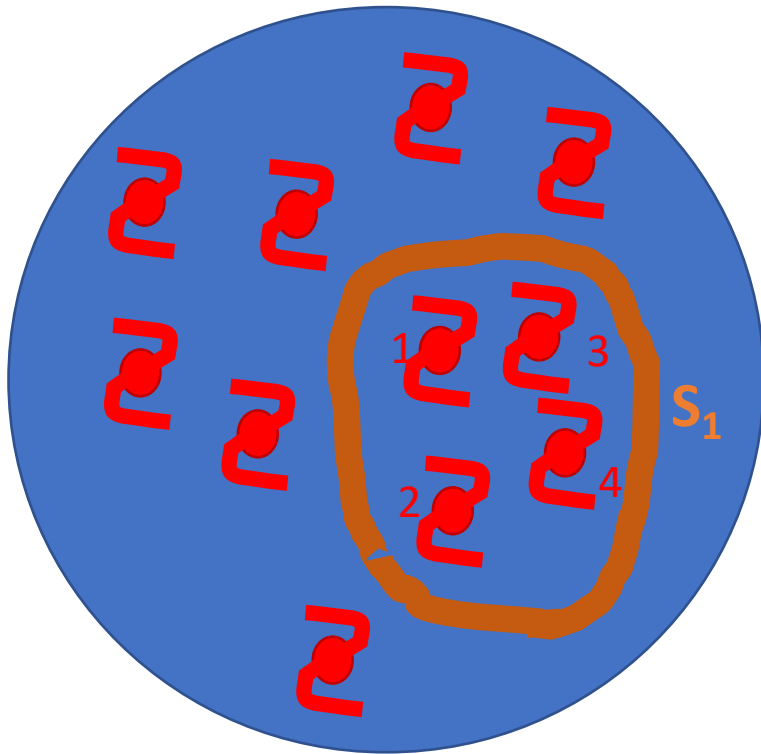
$$S_1 = \{x_1, x_2, x_3, x_4\}$$

$$S_1 = \{19.45, 16.20, 16.43, 19.10\}$$

$$\hat{\mu}_1 = 17.8$$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



$$S_1 = \{x_1, x_2, x_3, x_4\}$$

$$S_1 = \{19.45, 16.20, 16.43, 19.10\}$$

$$\hat{\mu}_1 = 17.8$$

95.5 % CI on $\hat{\mu}_1$?

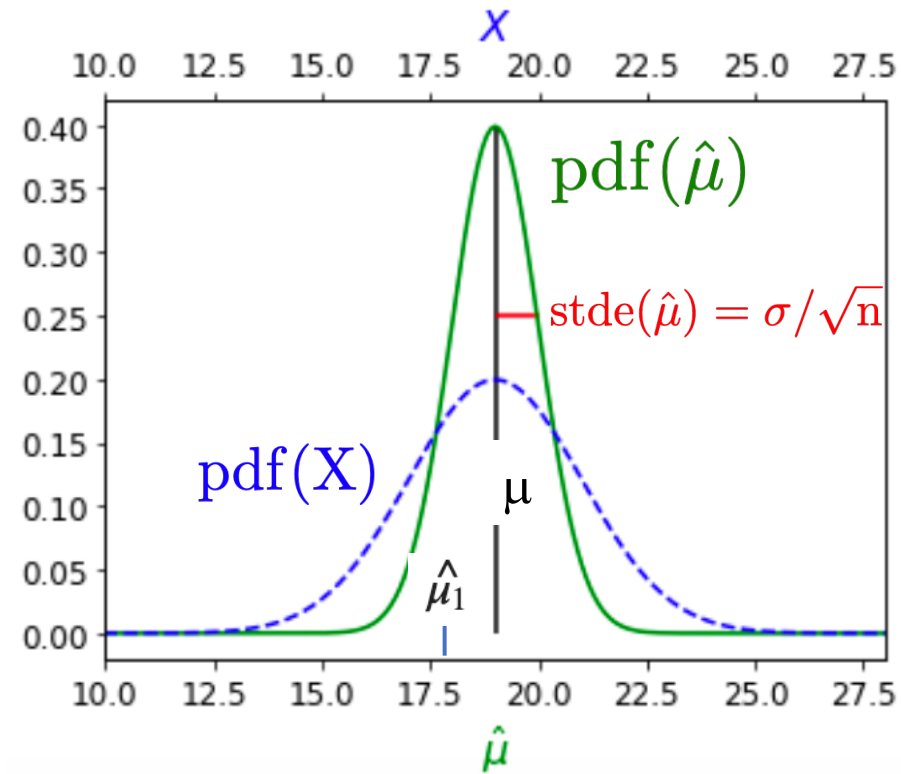
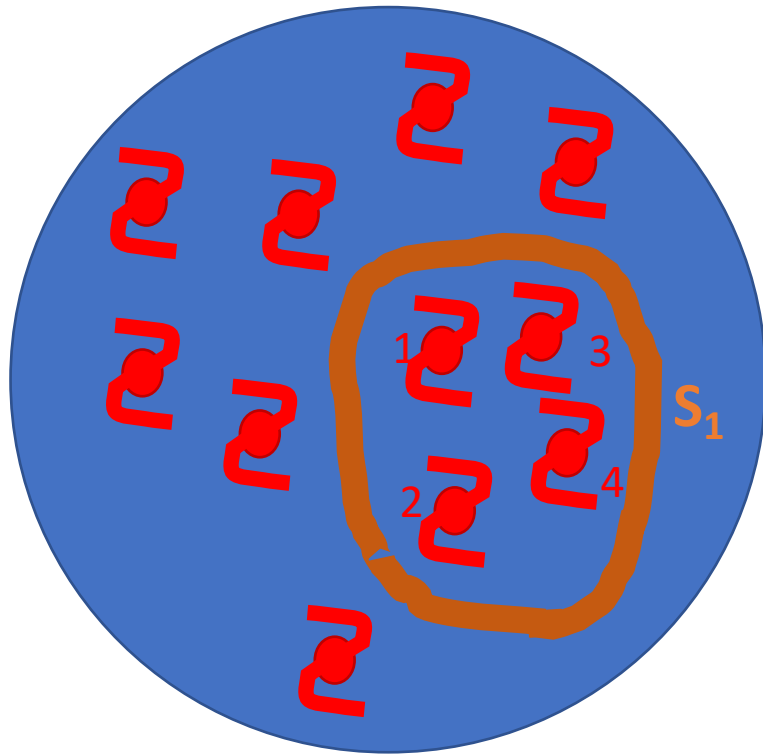


We need to look at the distribution of the estimator $\hat{\mu}$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$

95.5 % CI on $\hat{\mu}_1$?



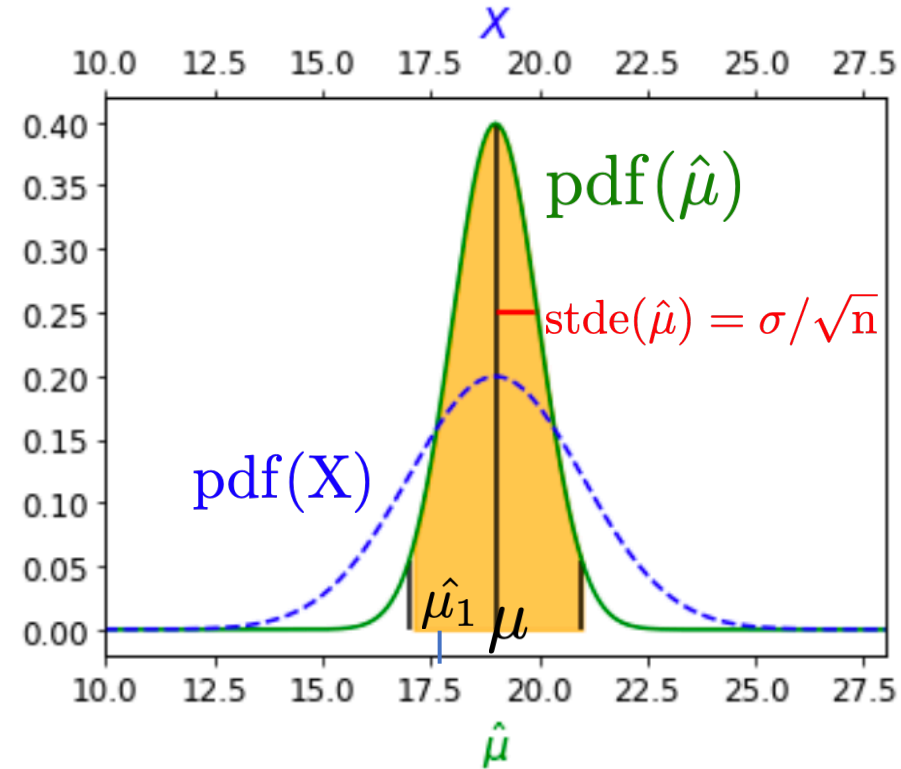
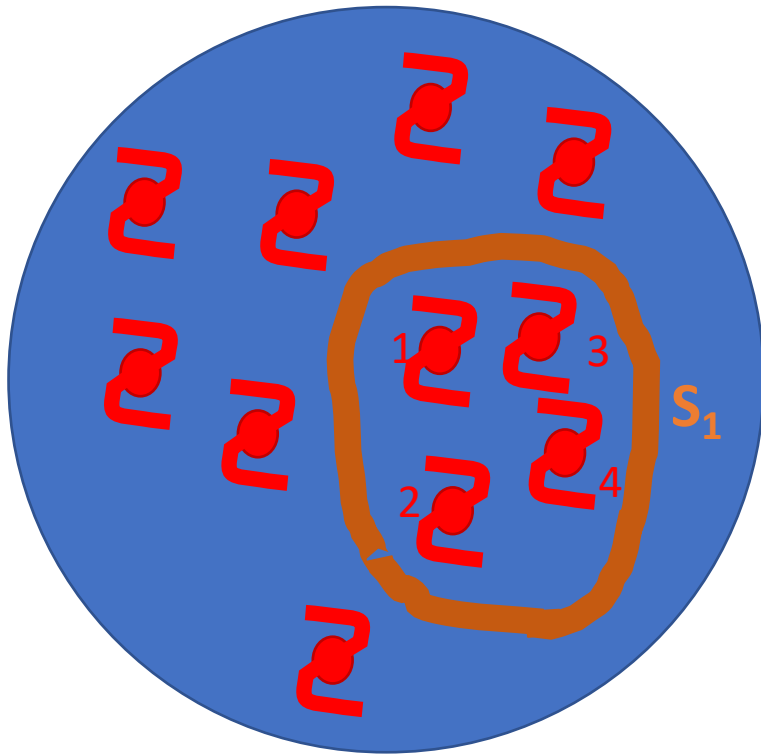
Case 1: σ is known

$$\hat{\mu} \sim N(\mu, \sigma/\sqrt{n})$$

$$\text{stde}(\hat{\mu}) = \sigma/\sqrt{n}$$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



95.5 % CI on $\hat{\mu}_1$?

Case 1: σ is known

$$\hat{\mu} \sim N(\mu, \sigma/\sqrt{n})$$

$$\text{stde}(\hat{\mu}) = \sigma/\sqrt{n}$$

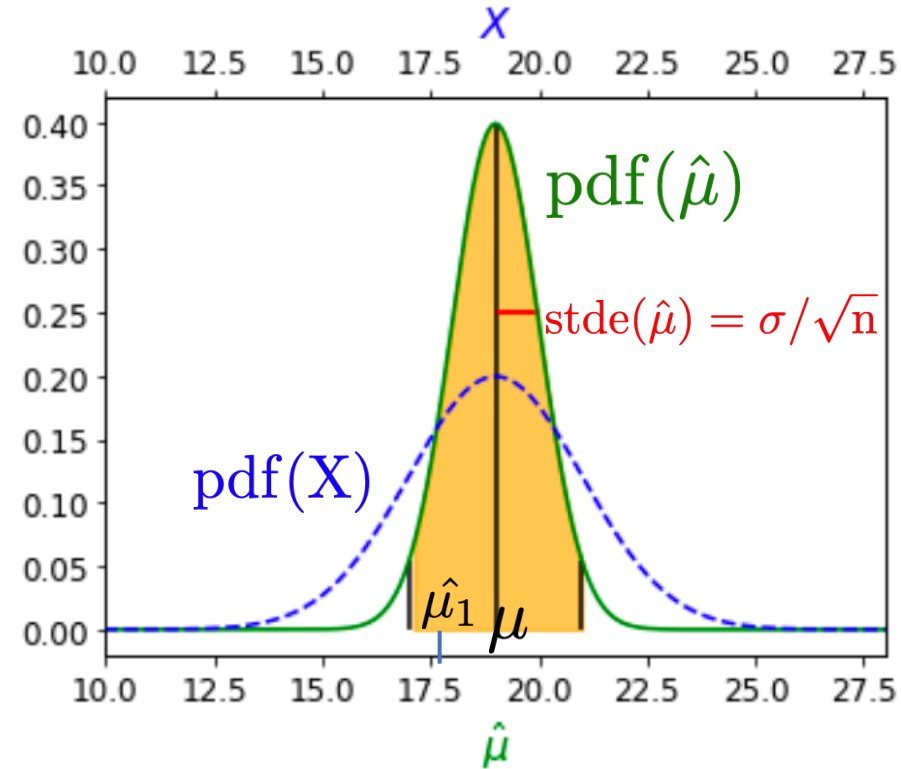
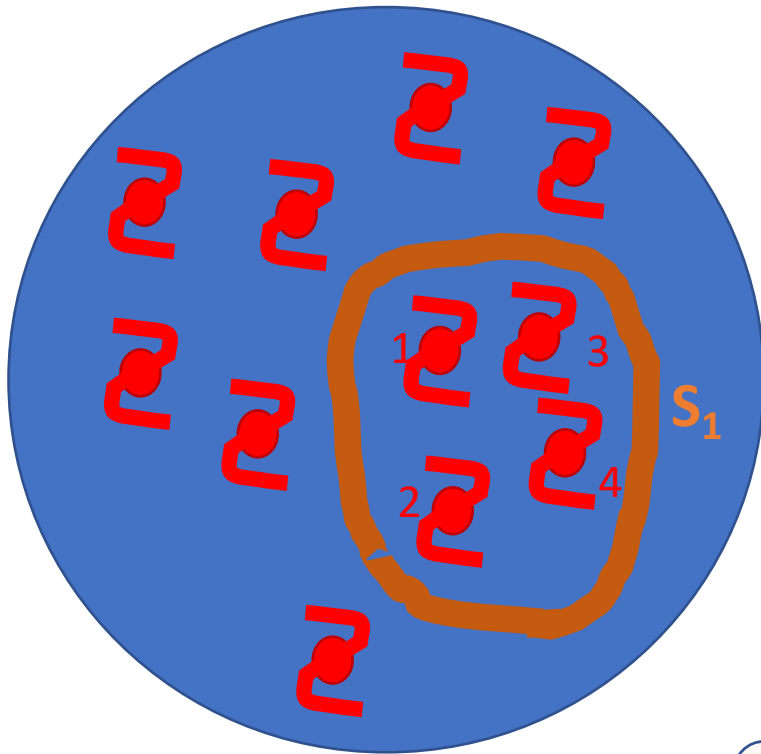
$$p(\hat{\mu}_1 \in [\mu \pm 2 \text{stde}]) = 0.955$$



$$p(\mu \in [\hat{\mu}_1 \pm 2 \text{stde}]) = 0.955$$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



$$\text{CI} \equiv [\hat{\mu} \pm m \times \text{stde}]$$

95.5 % CI on $\hat{\mu}_1$?

Case 1: σ is known

$$\hat{\mu} \sim N(\mu, \sigma/\sqrt{n})$$

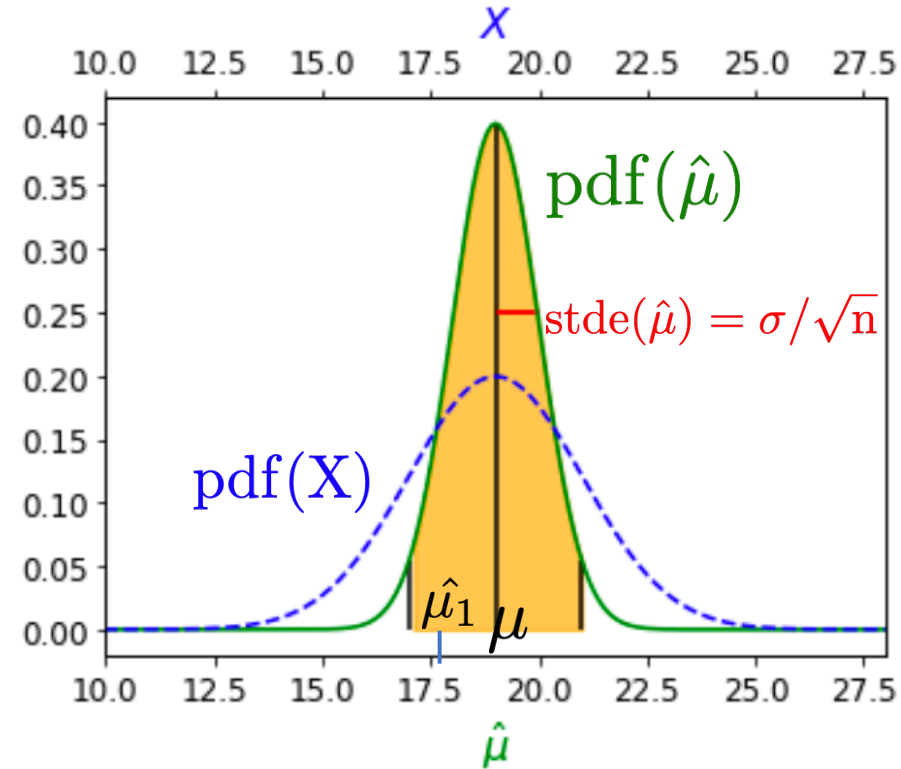
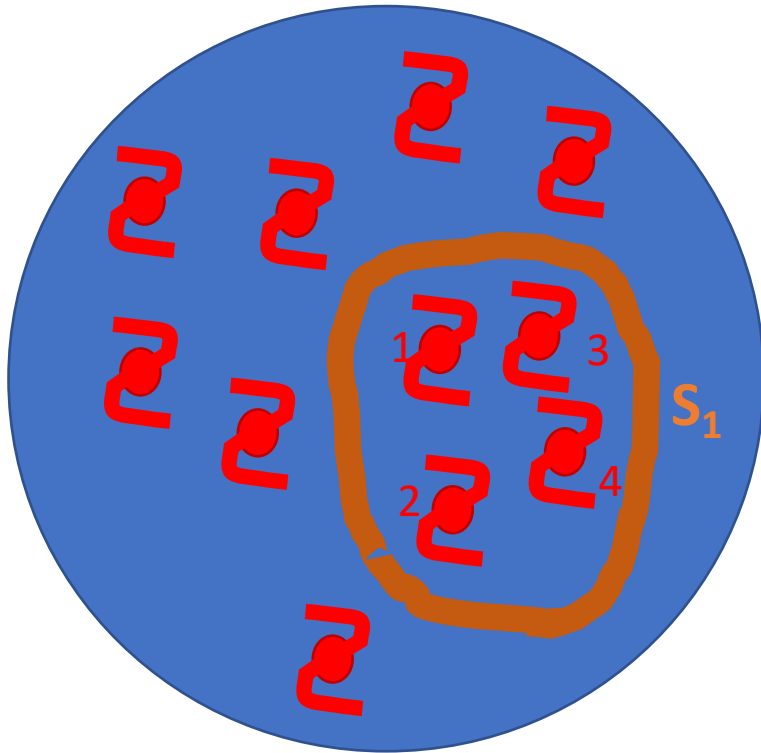
$$\text{stde}(\hat{\mu}) = \sigma/\sqrt{n}$$

m depends on the distribution of the normalised estimator:

$$\frac{\mu - \hat{\mu}}{\text{stde}}$$

Confidence interval on $\hat{\mu}$

If RV $\{X_i\}$ whose population is distributed as $N(\mu, \sigma)$



$$\text{CI} \equiv [\hat{\mu} \pm m \times \text{stde}]$$

95.5 % CI on $\hat{\mu}_1$?

Case 1: σ is known

$$\text{stde}(\hat{\mu}) = \sigma/\sqrt{n}$$

$$Z = \frac{\mu - \hat{\mu}}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Case 2: σ is unknown

$$\text{stde}(\hat{\mu}) = s/\sqrt{n}$$

$$t = \frac{\mu - \hat{\mu}}{s/\sqrt{n}} \sim t(n - 1)$$

Go to: [Student_vs_Gauss.ipynb](#)

Confidence interval

“Generic” strategy:

- Look at the PDF of the estimator of interest (the latter is often an output of Bayesian analysis) or of a normalised estimator of known distribution.
- Define region around your estimator (or a normalised estimator for which you know the distribution) that enclose **(1 - α) X 100 %** of the area under the PDF.

- For CI around the mean: $CI_{\alpha} = [\bar{x} - q_{\alpha/2} \text{ stde}, \bar{x} + q_{\alpha/2} \text{ stde}]$

$$q_{\alpha/2} = CDF^{-1}(1 - \alpha/2)$$

For CI on other statistics (e.g. variance, difference between 2 means, proportions) : see [Frequentist inference 03.ipynb](#)
(Material in [Frequentist inference 03.ipynb](#) is supplementary material)

Go to: Sect. II.2. of the notebooks