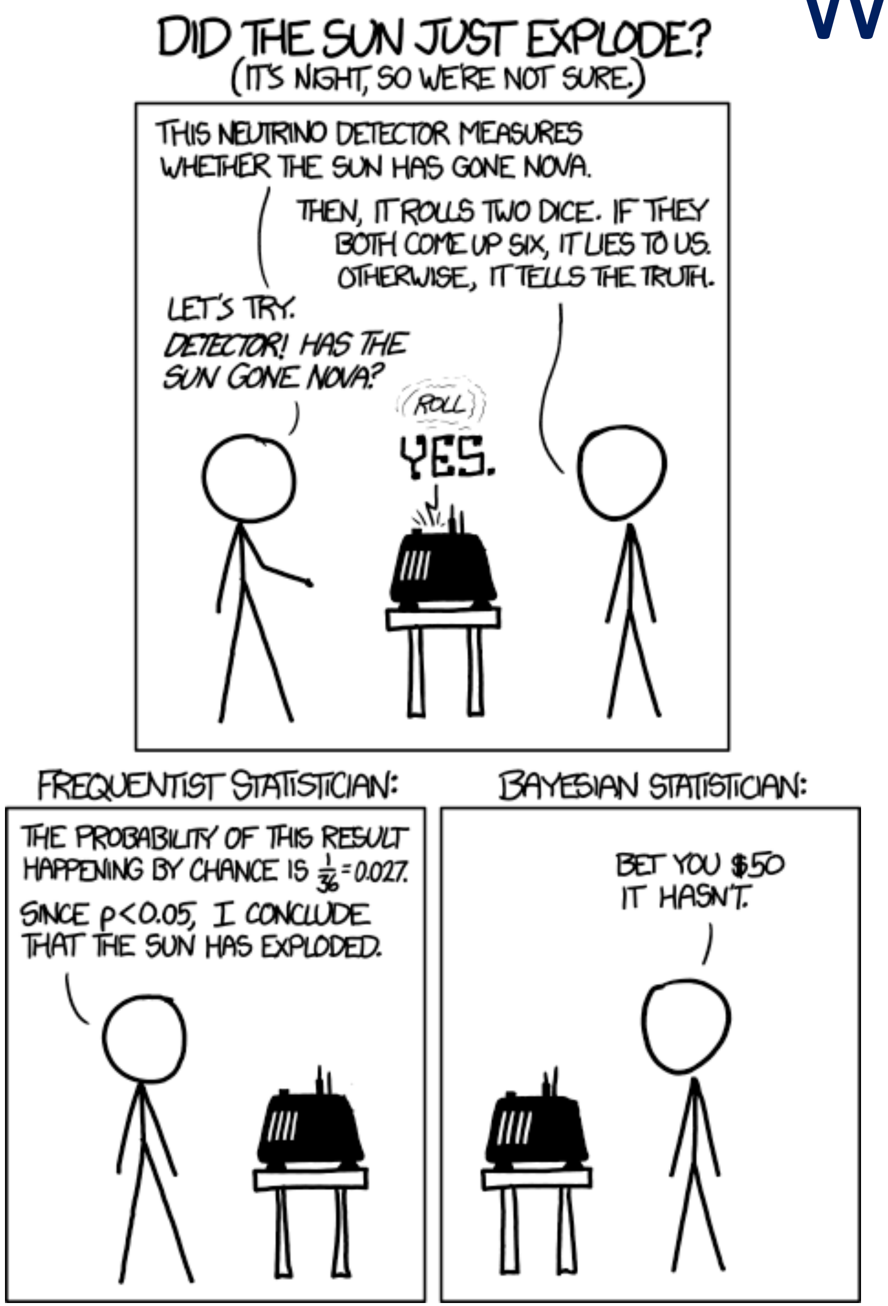
Classical statistical inference

Part 2

Associated notebook:

04-Basic_statistical_inference_frequentists_2/Frequentist_inference_01.ipynb

What is inference?



https://xkcd.com/1132/

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

Examples:

- Exoplanet transit ⇒ M, d ⇒ P(M | d)
- Supernovae distances ⇒ Expansion rate H_0

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

Three types of inference:

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

Point estimate $\hat{\theta}$

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)

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

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

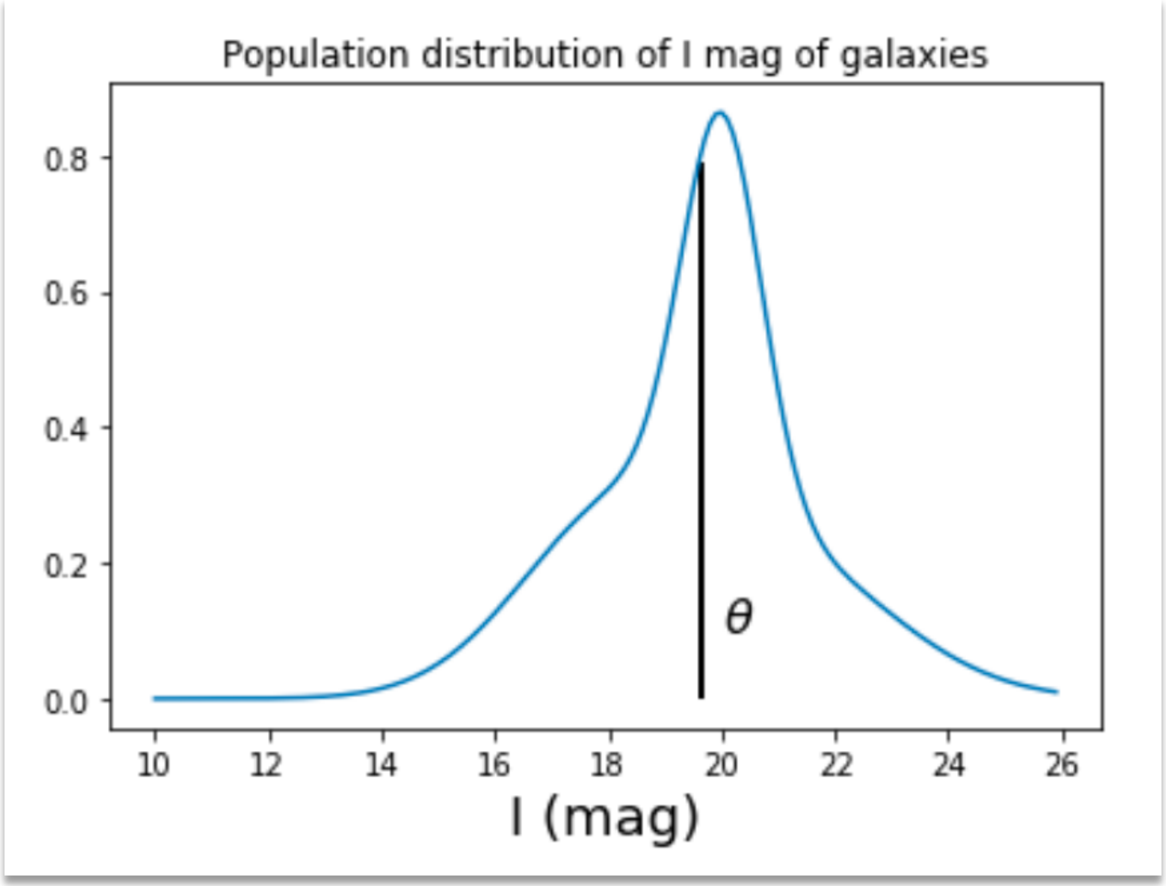
Point estimate $\hat{\theta}$

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, …
- 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*

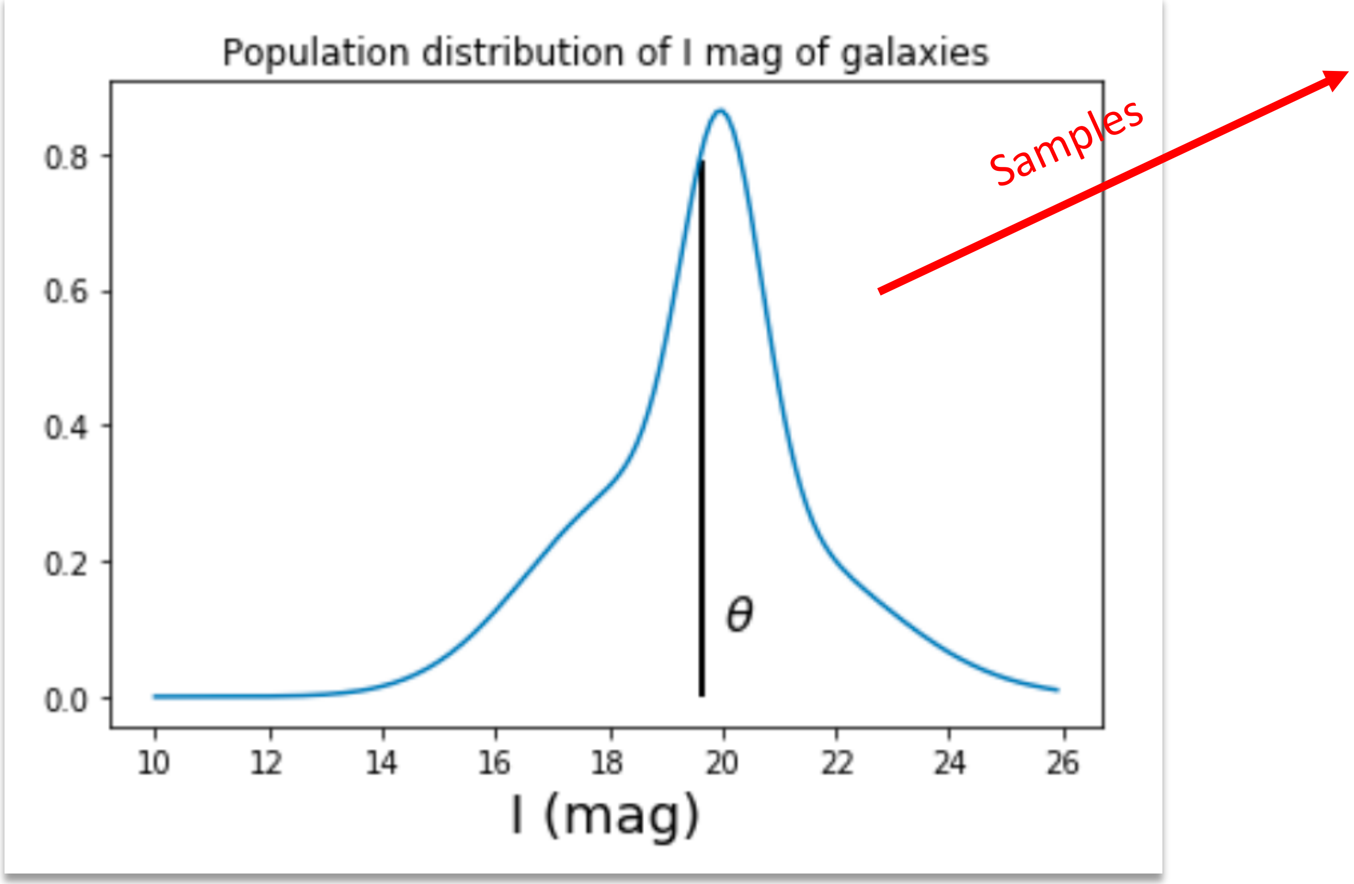
Point estimate $\hat{\theta}$

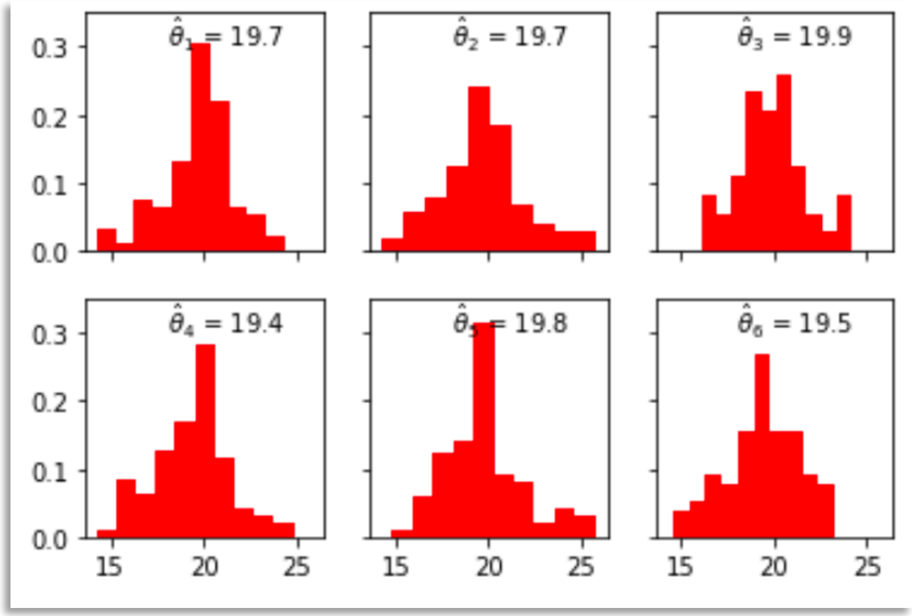


Population mean: $\theta = 19.66$

Point estimate $\hat{\theta}$

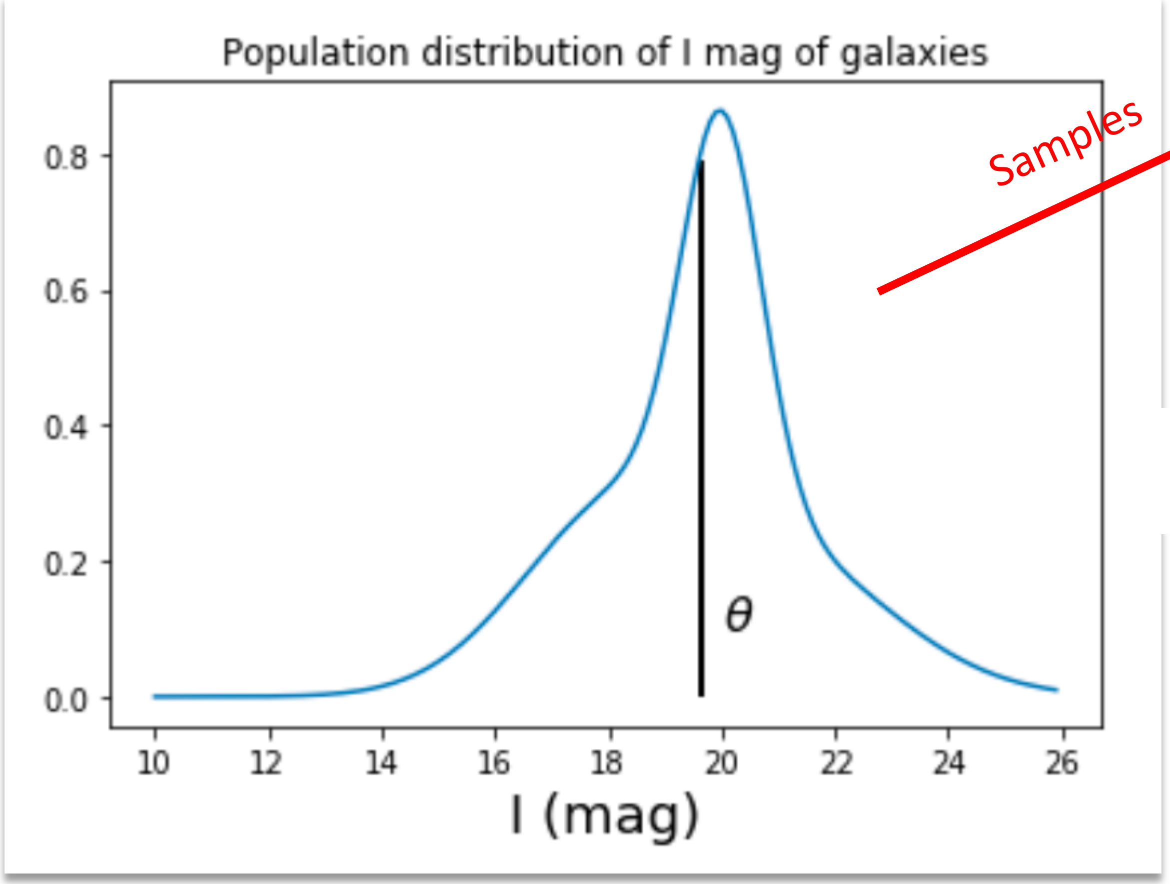
Six different *samples* drawn from the true population





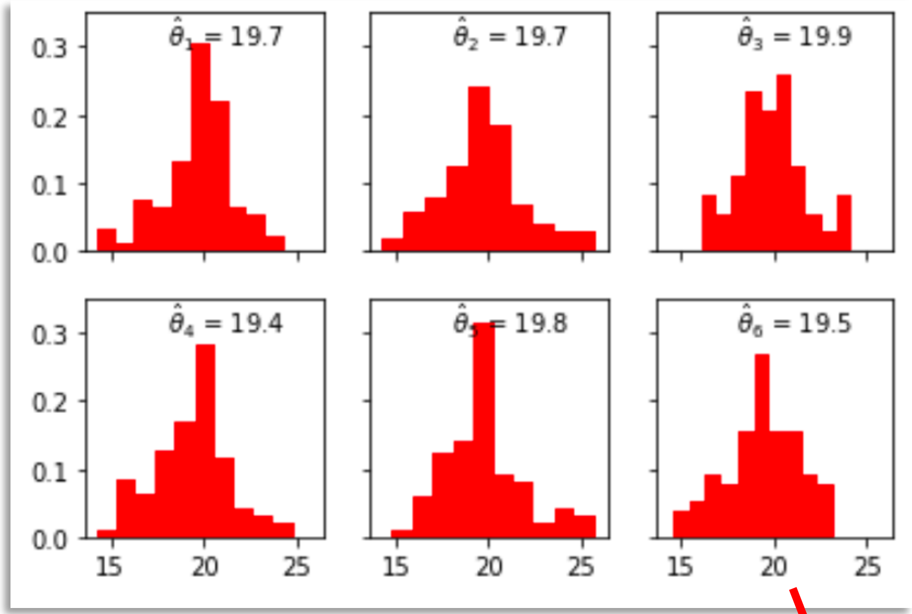
Population mean: $\theta = 19.66$

Point estimate $\hat{\theta}$

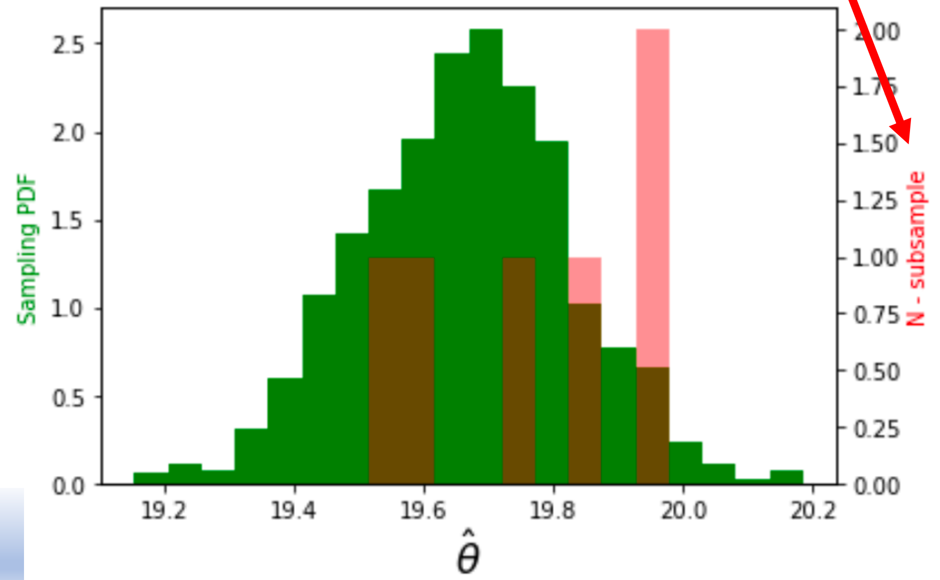


Samples

Six different *samples* drawn from the true population



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

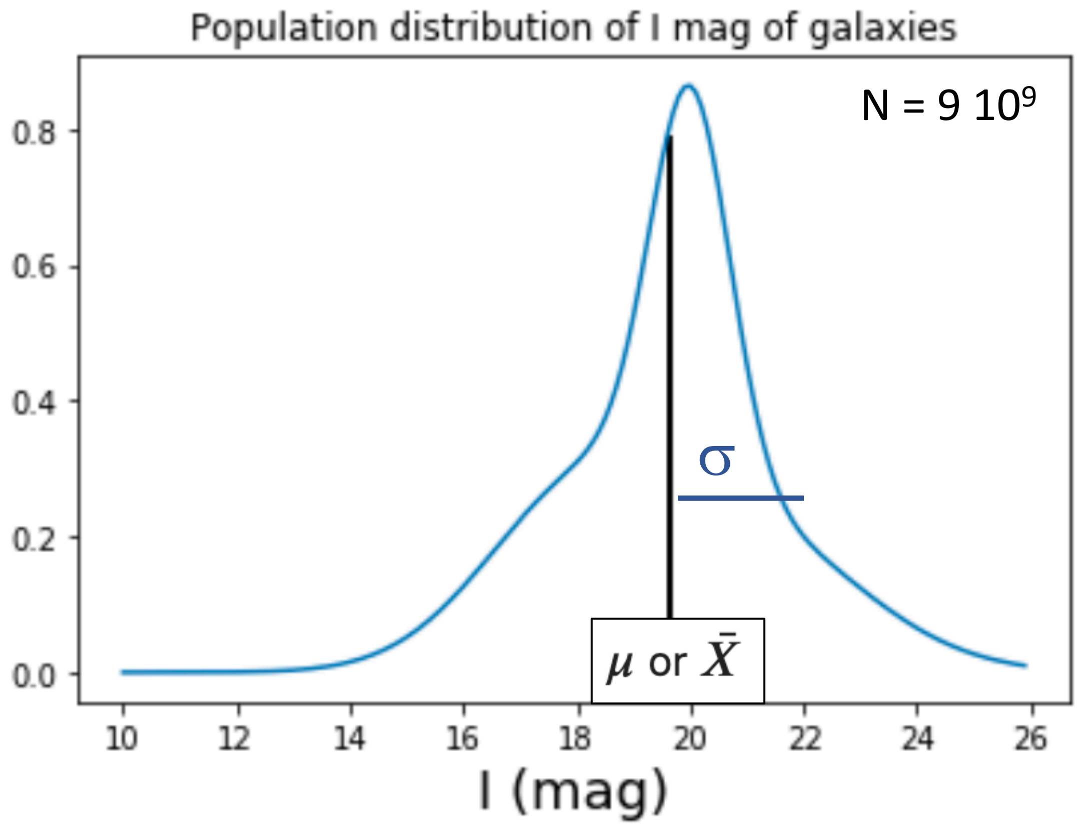


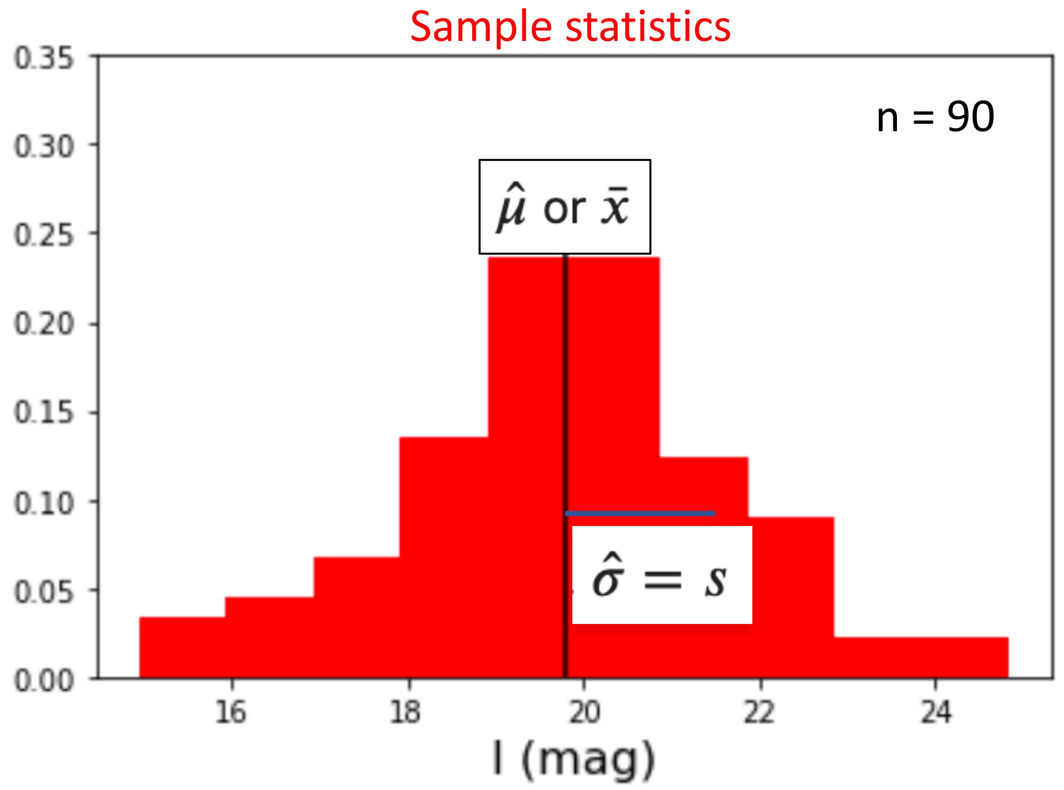
Population mean: $\theta = 19.66$

Go to: Sect. II.1 of the notebook

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	$\sigma = \sqrt{\sigma^2}$	$\hat{\sigma} = s = \sqrt{s^2}$





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 >>)

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

Student distribution

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

Chi square distribution