Recap

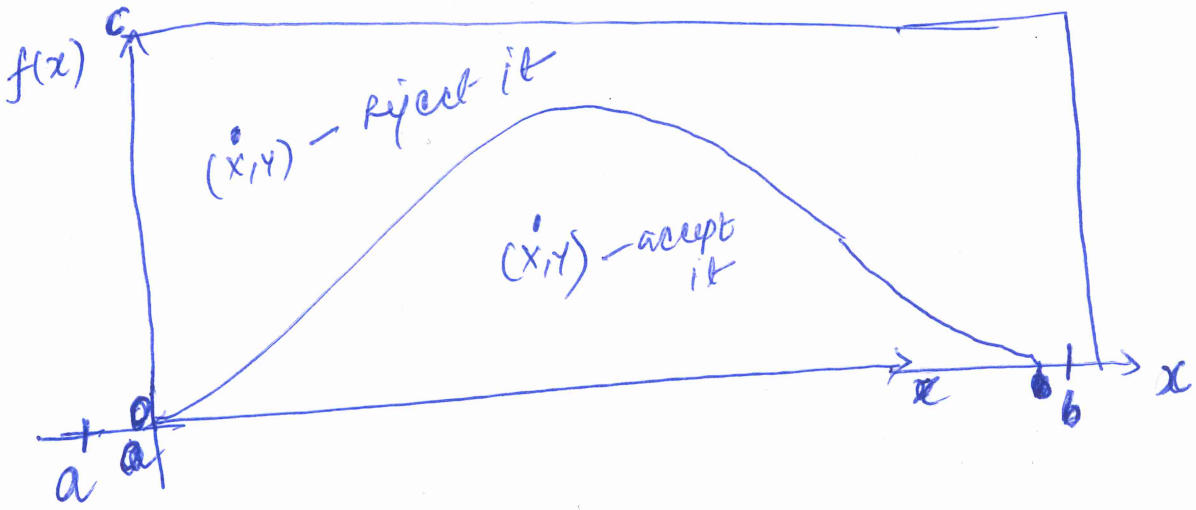
- Monte Carlo simulation can be used to compute (or estimate or approximate) features of the prob. distn. of X, e.g.,
  - $E[X]$, $E[g(X)]$
  - $\mathrm{Var}[X]$
  - $P[X \in A]$ — Also saw how to choose N to achieve a desired level of accuracy (N = # draws)
  - •

- Monte Carlo integration

Rejection method for simulation from a continuous distribution

Suppose X has pdf f and cdf F. It is difficult to invert F to get $X = F^{-1}(U)$, but f has a simple form.

Result: Choose a point (X, Y) at random from under the graph of density f. Then X has pdf f.



How to simulate such a point?

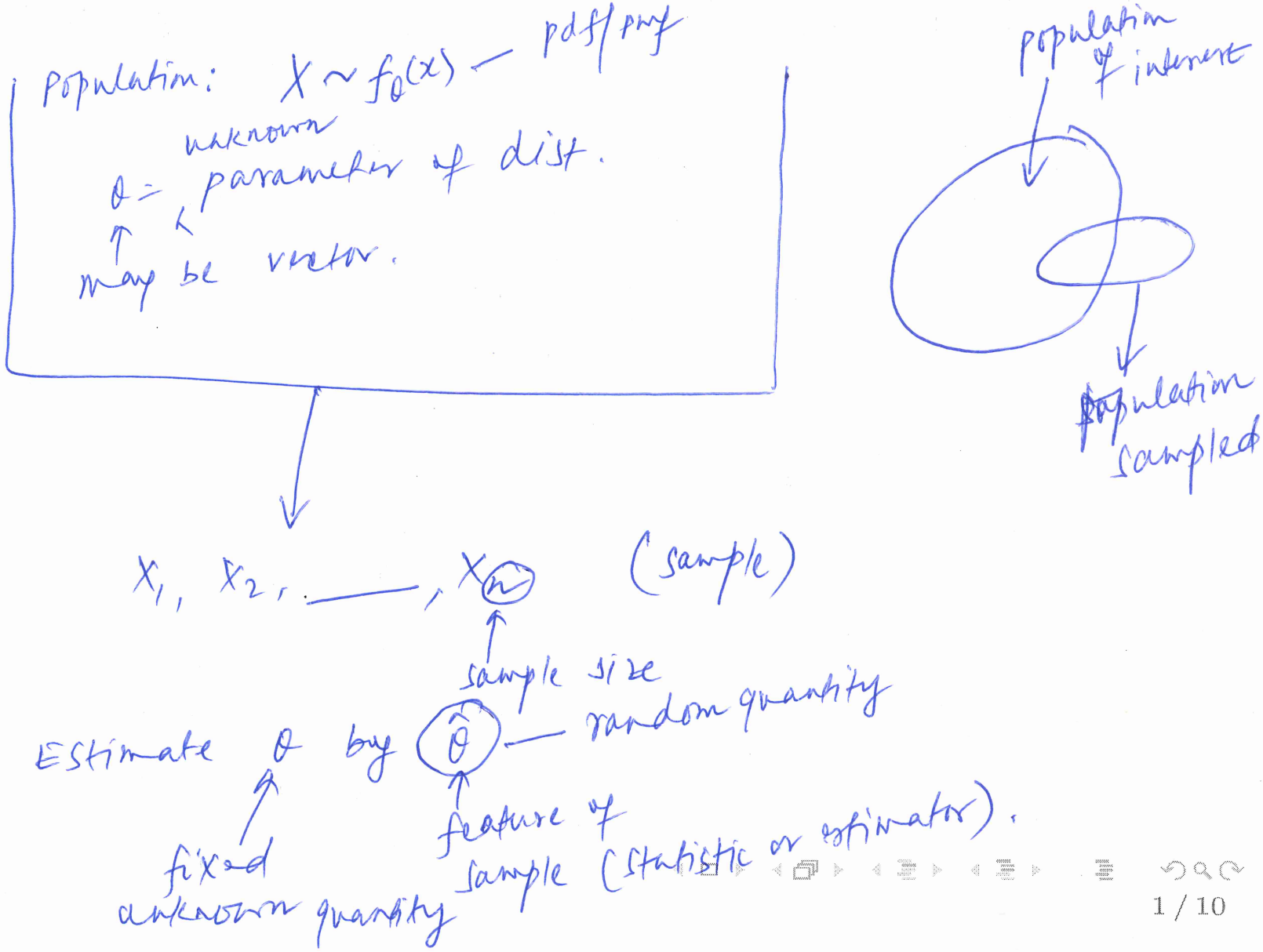
- Find a *bounding box* around the density curve, i.e., find a, b and c such that $a \le X \le b$, and $0 \le f(x) \le c$.
- Simulate a point (X, Y) *uniformly* distributed in the bounding box, i.e., simulate $X \sim \text{Uniform}(a, b)$, and $Y \sim \text{Uniform}(0, c)$.
- Reject the point if it falls above the density curve, i.e., $Y > f(X)$ and return to the previous step; otherwise accept the point, i.e., $Y \le f(X)$..

NOTE: This algorithm also works if we have a $c(x)$ such that $f(x) \le c(x)$ for all x.

- See book for an example of rejection method
- See R Handout for an example of HOW to do this in R.

Introduction to Statistics (Chapter 8)

Statistics: Learning about a population based on a sample from it. Recall a general *statistical inference* framework:



Statistic: Any feature of the sample data. They are used construct *estimators* of features of the population.

Sampling and non-sampling errors: Discrepancy between a sample and the whole population.

- *Sampling error* is caused by the fact that only a portion of the population is sampled. In most cases, this error reduces as n increases.
- *Non-sampling error* occurs if the sample is biased, i.e., it is not representative of the population of interest. Avoid well-known problems, such as selection bias, non-response bias, investigator bias, etc., while collecting data.

→ Assuming: there is no non-sampling error.

Random sample: $X_1, \ldots, X_n$ are independent and have the same distribution as X

- IID (independently and identically distributed) data
- Sample is representative of population.

BOOK.

Ex: To evaluate effectiveness of a processor for a certain type of tasks, we recorded the CPU time for $n = 30$ randomly chosen jobs (in seconds): 70, 36, 43, 69, 82, 48, 34, 62, 35, 15, 59, 139, 46, 37, 42, 30, 55, 56, 36, 82, 38, 89, 54, 25, 35, 24, 22, 9, 56, 19.
What is population? X? Sample? Distribution of X?

sample.

Population: CPU times of all jobs of the given type.

$X :=$ CPU time of a randomly selected ("typical") job.

Population: $X \sim f_\theta(x)$

Parametric analysis: Assuming a specific form for $f_\theta(x)$ (e.g., a gamma dist.).

Non-parametric analysis: No need to make strong assumptions about the shape of $f_\theta(x)$.

Desirable properties of an estimator $\hat{\theta}$ of θ

$\hat{\theta}$ will have a *probability distribution* — induced by randomness in the sampling process. It is called *sampling distribution* of $\hat{\theta}$.

Unbiasedness: (Repeated sampling justification)

- $\hat{\theta}$ is unbiased for θ if $E(\hat{\theta}) = \theta$ for all θ.
- Estimator is correct on average. In the long-run.

Small variance

- Variance = uncertainty.
- Larger variance = less precise.
- We would like to have small variance or high precision.
- Standard error (se) of $\hat{\theta}$ = standard deviation of $\hat{\theta}$

Player 1:

unbiased player,
but
high variability

average throw
= target

Player 2:

biased, but low variability

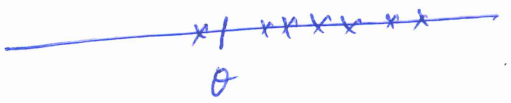
Player 3:

unbiased + low variability

Estimation:

unbiased

biased



unbiased + small variability

Consistency:

- $\hat{\theta}$ is consistent for θ if it converges to θ as $n \to \infty$.
- Necessary for a reasonable estimator.
- Why use an estimator that does not become more accurate as n increases?

Asymptotic normality:

- For large n, $\hat{\theta}$ approximately follows $N(\theta, \text{var}(\hat{\theta}))$.
- Consequence of CLT and related results.
- Useful for designing inference procedures that are valid for large n

Some descriptive statistics and what they estimate

Have a random sample $X_1, \ldots, X_n$ from the population of X.

Mean:

Population mean: $\mu = E[X]$

Sample mean: $\overline{X} = \sum_{i=1}^{n} X_i / n$

Properties of $\overline{X}$: natural

- $\overline{X}$ is a(n) estimator of μ [may be possible to find better estimator]
- LLN: $\overline{X}$ is consistent.
- $E[\overline{X}] = \mu$ for all $\mu \Rightarrow \overline{X}$ is unbiased. $\mathrm{Var}[\overline{X}] = \frac{\sigma^2}{n}$ $\mathrm{Var}(X)$
- CLT: $\overline{X} \dot{\sim} N[\mu, \frac{\sigma^2}{n}]$ if n is large (Asymptotic normality)
- Greatly affected by outliers

Ex: (CPU data): $\overline{X}$=?