in a paper in 1934 that laid the foundations (51)
for contemporary scientific sampling. (17 Jan 19)

Two simple ways to do random sampling: at random with replacement (independent identically distributed (IID) sampling) & at random without replacement (simple random sampling (SRS)).

Toy example

population   sample

$$N = 3 \begin{bmatrix} 1 \\ 2 \\ 7 \end{bmatrix} \xrightarrow{\text{at random}} \begin{matrix} I_1 \\ I_2 \end{matrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \quad n = 2$$

population size

sample size

capital

Let I_1, I_2 be random variables (AMS 131) recording the results (y_1, y_2) of $n = 2$ random draws from this toy population;

lower case

B_i = (B_i), $i = 1, 2$; B_1 = (draws are IID); B_2 = (draws are SRS)

$P(I_1 = 9 \overset{\text{and}}{,} I_2 = 9 \mid \overset{(\text{IID})}{B_1}) = \frac{1}{9}$ but independent draws

$P(I_1 = 7, I_2 = 7 \mid \overset{(\text{SRS})}{B_2}) = 0$ ← dependent draws

Facts about IID sampling & SRS

① Both methods promote but (of course) do not always achieve similarity of the sample $\mathcal{S}$ & the unsample $\mathcal{U}$.

Def. A sampling method is strong-sense representative if all possible subsets of size n from a population of size $N \ge n$ have the same probability.

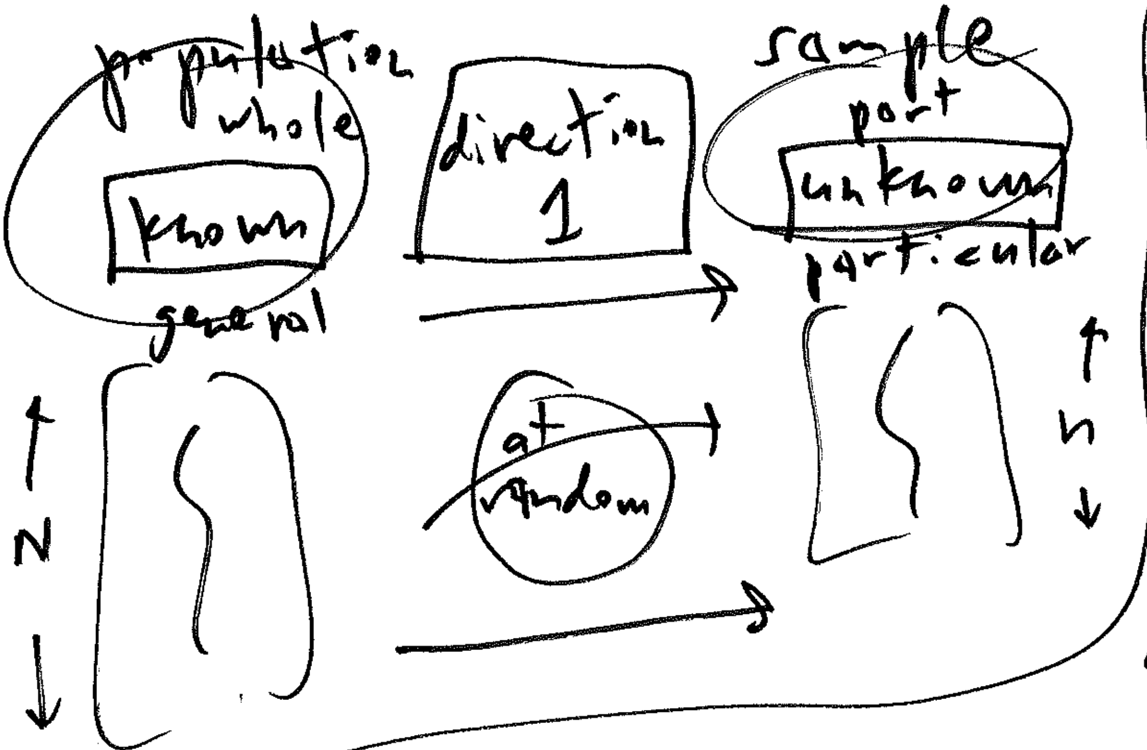
IID ($\frac{1}{9}$ each)	SRS ($\frac{1}{6}$ each)
11	12
12	17
17	21
21	27
22	71
27	72
71	
72	
77	

② IID and SRS are both strong-sense representative.

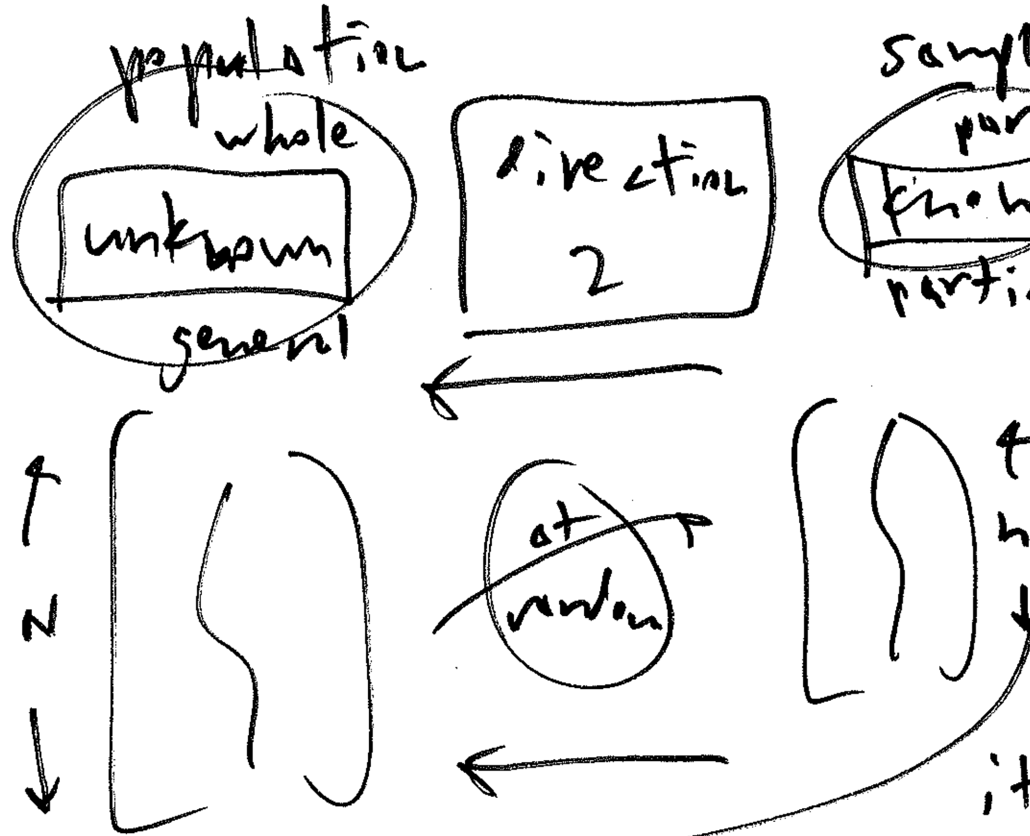
③ The idea behind representative sampling is as follows: Ⓐ A complete census of all N population individuals is typically at best too expensive in (time, money) & at worst impossible (example of impossibility: sample = {all eBay users who clicked on at least one eBay web page between (1 Mar - 31 May 2018)}; unsample, contemplated on 1 Apr 2018, = {all eBay users who will click on at least one eBay web page between (1 Jul - 30 Sep 2018)}.

(*) next page

(B) If the sampling method is representative, you can use the observed data in the sample to estimate the unobserved data in the unsample & thereby draw inferences about the entire population P. The population & sample diagram can be "run" in 2 different directions:



This is **deductive** reasoning; when the sample is drawn randomly it's an application of **probability**



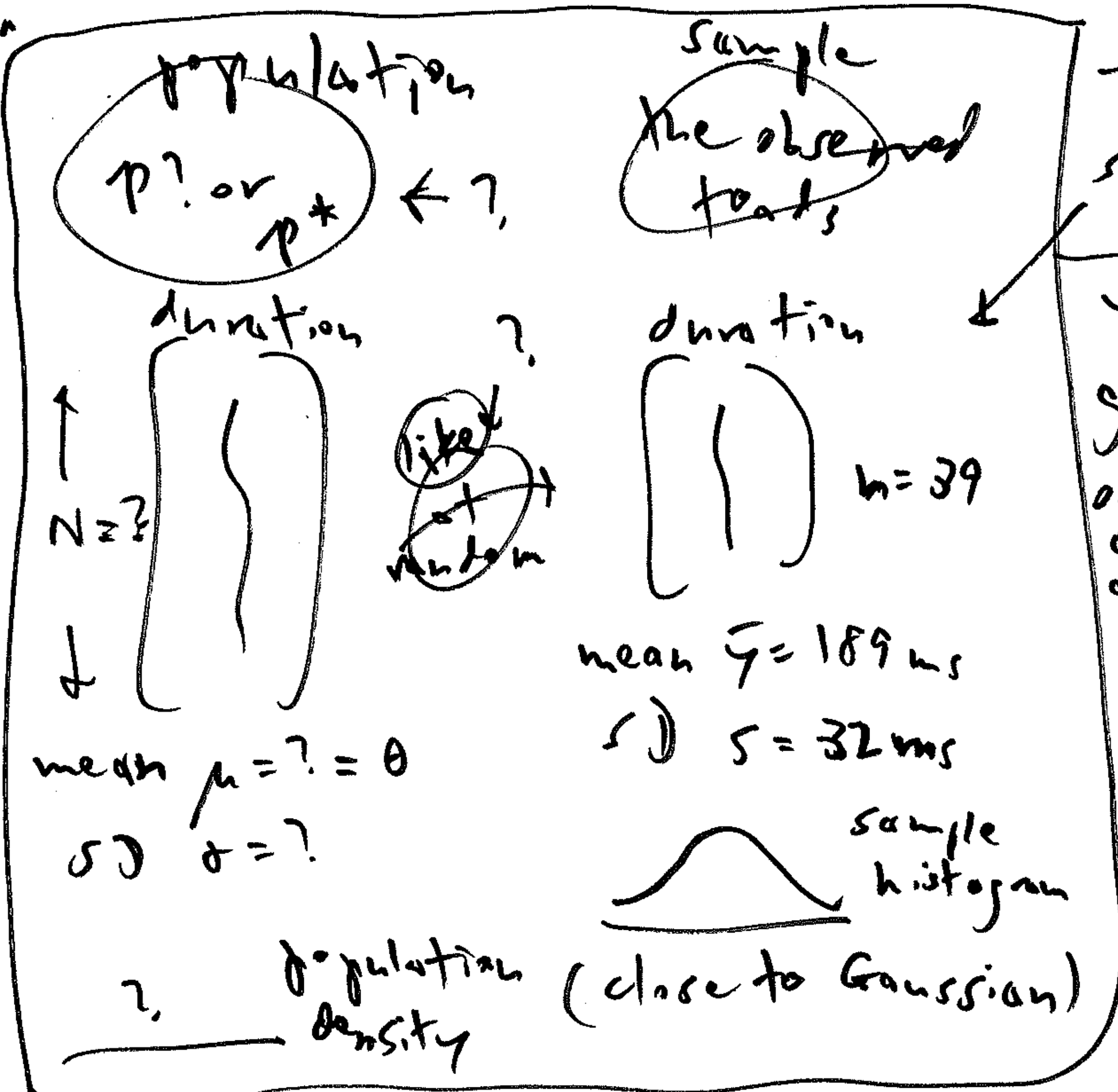
This is **inductive** reasoning; when the sample is drawn randomly it's an application of (**statistical**) **inference**.

⊛ Important practical note:

random sampling is a goal that's not always achievable in practice.

Example: (AMS 7)

A biologist named Dr. Littlejohn measured the duration of mating calls of $n = 39$ Tasmanian male tree toads (1 call per toad) in 1965; this formed his sample. He wanted to generalize outward to an appropriate population of tree toads. He made a serious effort to achieve a sample



that was representative of (like a random sample from) P = {all male tree toads of this species in Tasmania in 1985}. It's clearly impossible to literally take a random sample from $\mathcal{P}$. One safe statement can always be made: his 39 toads are like a random sample from $\mathcal{P}^* = \{$all male tree toads of this species similar to his toads in all relevant ways$\}$.

mating call duration

More facts about IID sampling & SRS

③ if $n = 1$, SRS = IID (no chance yet to make a 2nd draw)

④ if $n \ll N$, SRS $\doteq$ IID (in that case, P(sampling the same individual more than once | $\mathcal{B}$) = tiny

is a lot smaller than

IID

⑤ the math is easier for IID (independence) but people actually do SRS in practice;

when $n << N$ (often true), this supports (57)
the common practice of (sampling with SRS but analyzing your data with IID formulas).

Back to CS2 | we want to do inference outward from the $n = 403$ AMI patients at DH from 2015-2018. Inference is inherently probabilistic.

The Pascal-Fermat (P-F) (classical) approach to probability, if attempted directly would require us to consider each possible random sample as an EO; with e.g. $N = 1,000$ patients in the population, there would be $\binom{1000}{403} \doteq 1.6 \cdot 10^{291}$ such samples (impractical).

History of probability, continued

~~Cardano (1520s)~~

Pascal - Fermat (1654) / (1713) ((James) Jacob Bernoulli)
weak law of large Numbers for binary outcomes (1733) (Abraham de Moivre)

Central Limit Theorem (CLT) for binary outcomes.

(1763) (Thomas Bayes) used conditional probability to draw statistical inferences (learn about population summaries from sample data), also with binary outcomes

(1774) (1812) (1814) (Pierre Simon de Laplace) extended Bayesian reasoning to (many) other types of outcomes, including new CLTs

(1810) Laplace also developed the first <u>frequentist</u> probability reasoning (more below).

(1821) (Carl Friedrich Gauss) much influenced by Laplace; used Bayesian & frequentist methods to solve important problems in astronomy & other sciences.

(1837) (Siméon Denis Poisson) drew a distinction between "subjective" and "objective" probability

& disparaged the former (more on this later)

(~1837) (John Stuart Mill, Robert Leslie Ellis, Antoine Augustin Cournot) advocated frequentist probability over Bayesian reasoning; (1866) John Venn did the same in influential textbooks, as did George Boole (1854) & Joseph Bertrand (1889)

The frequentist story

Restrict attention to phenomena that are inherently repeatable under (essentially) identical conditions; let A = something of interest to you that either happens or doesn't happen in each repetition; then **Definition**

$$P_F(A) = \lim_{n \to \infty} \frac{\#\text{ times } A \text{ happens in first } n \text{ repetitions}}{n}$$