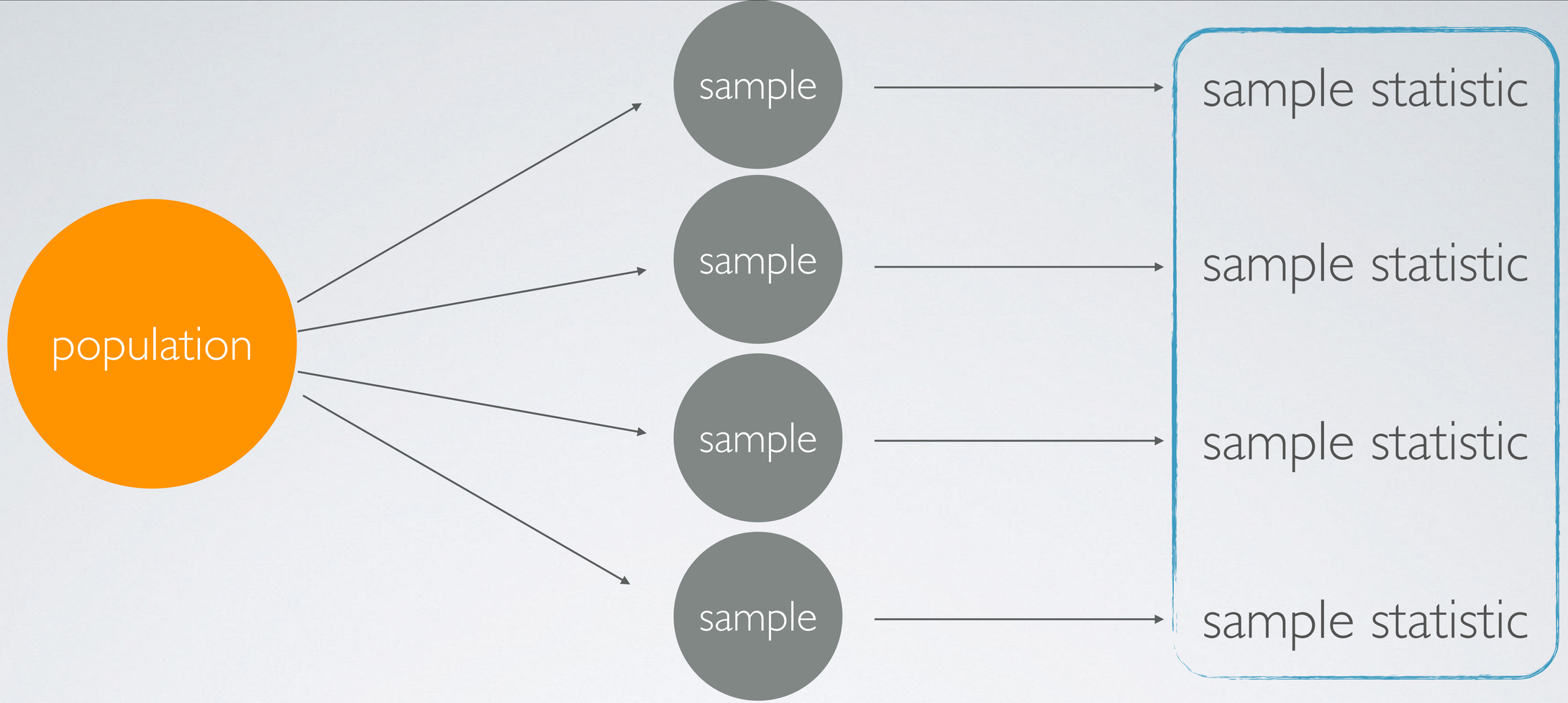
sampling variability & CLT for proportions

- sampling distribution
- CLT for proportions + conditions

Dr. Mine Çetinkaya-Rundel
Duke University

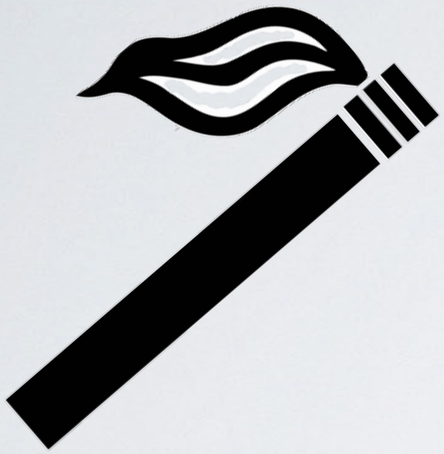


sample distributions ≠ sampling distribution

smoker or not
categorical

% smoker
numerical

%?



world population
N = pop size
p

Afghanistan: $x_{AF,1}, x_{AF,2}, \cdots, x_{AF,1000}$

...

USA: $x_{US,1}, x_{US,2}, \cdots, x_{US,1000}$

...

Zimbabwe: $x_{ZW,1}, x_{ZW,2}, \cdots, x_{ZW,1000}$

$\hat{p}_{AF}$

...

$\hat{p}_{US}$

...

$\hat{p}_{ZW}$

$$p = \frac{\#\text{ of smokers in the world}}{N}$$

sampling distribution

$$mean(\hat{p}) \approx p$$

CLT for proportions: The distribution of sample proportions is nearly normal, centered at the population proportion, and with a standard error inversely proportional to the sample size.

$$\hat{p} \sim N\left(mean = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

shape center spread

Conditions for the CLT:

1. ***Independence:*** Sampled observations must be independent.
   - random sample/assignment
   - if sampling without replacement, n < 10% of population
2. ***Sample size/skew:*** There should be at least 10 successes and 10 failures in the sample: $np \geq 10$ and $n(1-p) \geq 10$.

if p unknown, use $\hat{p}$

practice

90% of all plants species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants.

$p = 0.90$

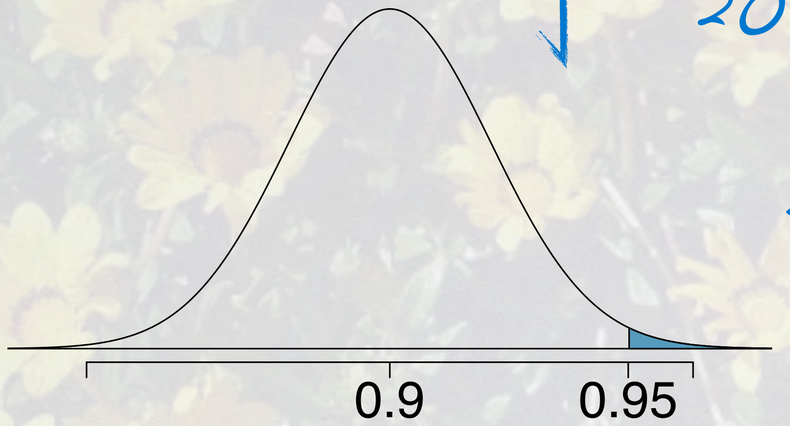
$n = 200$

$P(\hat{p} > 0.95) = ?$

1. random sample & <10% of all plants → independent obs.

2. $200 \times 0.90 = 180$ and $200 \times 0.10 = 20$

$\hat{p} \sim N(\text{mean} = 0.90, SE = \sqrt{\frac{0.90 \times 0.10}{200}} \approx 0.0212)$



$Z = \frac{0.95 - 0.90}{0.0212} = 2.36$

$P(Z > 2.36) \approx 0.0091$

practice

90% of all plants species are classified as angiosperms (flowering plants). If you were to randomly sample 200 plants from the list of all known plant species, what is the probability that at least 95% of plants in your sample will be flowering plants.

$p = 0.90$

$n = 200$

$P(\hat{p} > 0.95) = ?$

Using the binomial distribution:

$200 \times 0.95 = 190$

R

```
> sum(dbinom(190:200, 200, 0.90))
[1] 0.00807125
```

what if ?

if the success-failure condition is not met:

- the center of the sampling distribution will still be around the true population proportion
- the spread of the sampling distribution can still be approximated using the same formula for the standard error
- the shape of the distribution will depend on whether the true population proportion is closer to 0 or closer to 1

shape of the sampling distribution