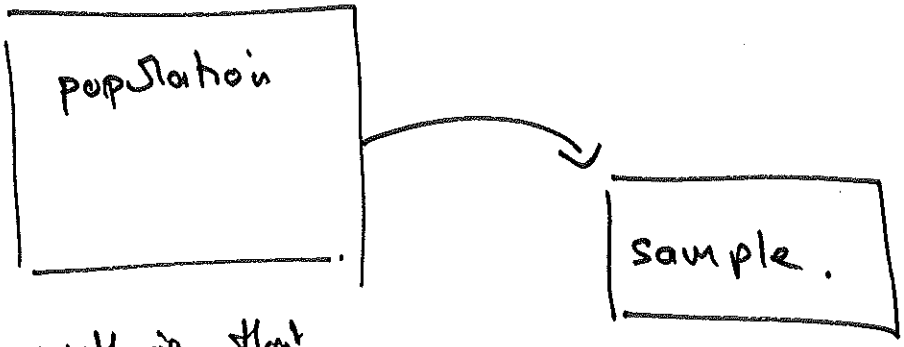



everything that you're interested in

the part of the population about which you have collected data.

Not measuring the entire population,
⇒ there will be <u>chance error</u> in the sample.

How big will this chance error be?
What does it depend on?

Somewhat unintuitively, - depends mostly on the size of the <u>sample</u>, much less on the size of the <u>population</u>.

Start by looking at the chance error in <u>percentage</u>.

(population is in different categories; look at the % of the sample that's in a particular category).

AMS 5 students

| | |
|---|---|
| 79 | fresh |
| 83 | soph. |
| 24 | juniors |
| 16 | seniors |

202 total.

consider taking a simple random sample of this population.

Sampling <u>without</u> replacement

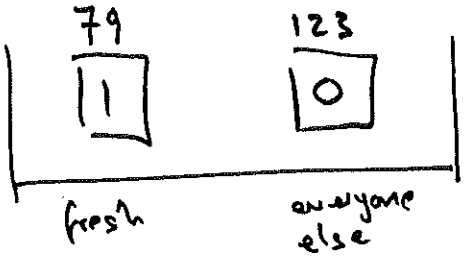
if sample size $\ll$ population size this is ok.

sample of size 10

Expected % fresh in our sample?

Size of the chance variation?

For a simple random sample, the expected value for the sample % equals the population %.




$$\text{mean of box} = \frac{79}{202} = 0.39$$

$$SD_{box} = (1-0)\sqrt{\frac{79}{202} \times \frac{123}{202}}$$

$$= 0.488.$$

Sample of size 10.

$$\text{Expected \# fresh} = 10 \times 0.39 = 3.9$$

$$SE = \sqrt{10} \times 0.488 \approx 1.5$$

expect $3.9 \pm 1.5$ fresh.

What about %?

Expected % of fresh. $= \frac{\text{expected \#}}{\text{sample size}} \times 100$

$= \frac{3.9}{10} \times 100 = 39\%.$

(≡ % in the population)

chance error. $= \frac{SE}{\text{sample size}} \times 100$

$= \frac{1.5}{10} \times 100 = 15\%.$

We expect the sample to be $39\% \pm 15\%$ fresh.

---

Take a sample of size 40.

Expected # fresh. $= 40 \times 0.39$

(sample size; mean of box)

Expected % fresh $= \frac{\text{Expected \#}}{\text{sample size}} \times 100.$

$= \frac{40 \times 0.39}{40} \times 100 = 39\%$

chance error.

$$\text{SE \# fresh} = \sqrt{40} \times 0.488.$$

$$\text{chance error in \% fresh} = \frac{\sqrt{40} \times 0.488}{40} \times 100$$

$$= \frac{0.488.}{\sqrt{40}} \times 100$$

$$= 7.7\%$$

Sample of size 40,

39 ± 7.7% fresh.

Sample of size 10

39 ± 15% fresh

Sample size has increased by a factor of 4.

chance error has decreased by a factor of 2.

As sample size increases

The chance error in % decreases.

↑

This is true irrespective of the size of the population, provided that the sample is small enough that we can consider the sample to be approximately drawn with replacement.

What happens when the sample is a significant fraction of the population?

The contents of the box get smaller on each draw ⇒ slightly less variability.

SE when drawing without replacement = SE when drawing with replacement × correction factor

$$\text{correction factor} = \sqrt{\frac{\#\text{ tickets in box} - \#\text{ draws}}{\#\text{ tickets} - 1}}$$

Sample of 10 from population of 202

$$\text{correction factor} = \sqrt{\frac{202 - 10}{202 - 1}} = 0.98$$

For a sample of 40 from a population of 202

$$\text{correction factor} = \sqrt{\frac{202 - 40}{202 - 1}} = 0.9$$

ie the chance error should be reduced from the value found earlier (7.7%) and should be

$7.7 \times 0.9 =$ ~~6.87~~ 6.9%

When sampling <u>without</u> replacement, the chance variability in % will be ~~also~~ smaller than the chance variability when sampling with replacement.

---

If we fix the <u>sample size</u>, how big must the <u>population</u> be before the correction is negligible?

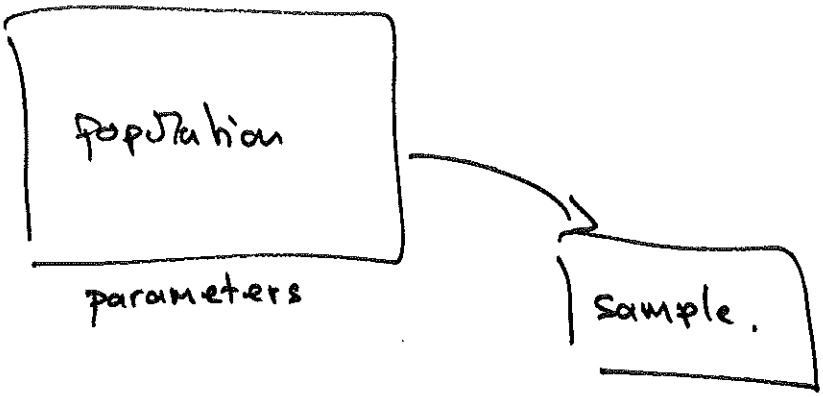
Opinion poll ~~of~~ with sample size $n = 2500$

| <u>population size</u> (50:50 split) | <u>correction factor</u> | |
|---|---|---|
| 5,000 | 0.707 | |
| 10,000 | 0.866 | |
| 100,000 | 0.987 | |
| 500,000 | 0.997 | ← really makes no difference. |

So far, we've assumed we've known what was in the box (population) and looked at the characteristics of a sample.

Now: turn it around — given a sample what can we say about the population?

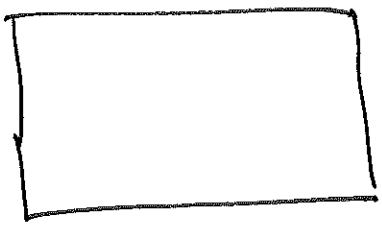
This is called inference.



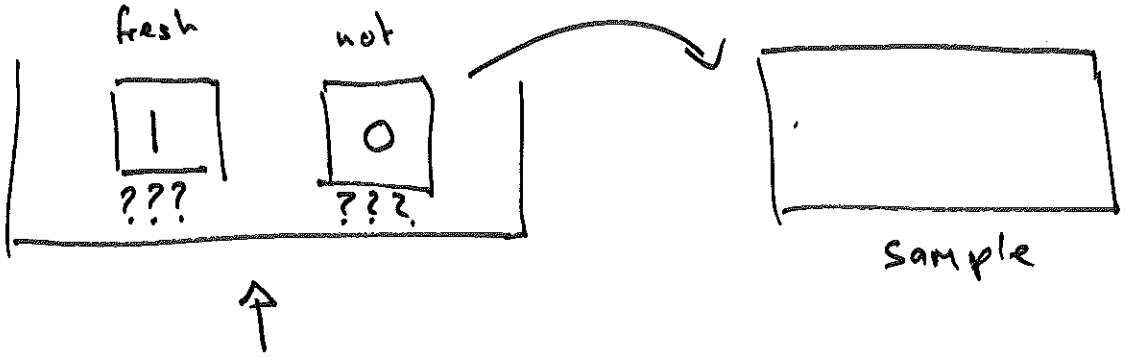

Assume SRS

from the size and composition of the sample, we can say how accurate the estimate of the population parameter will be.

Population: all UCSC students



want to know % fresh.




we don't know how many tickets of each type there are (this is what we want to find out!)

Assume we sample 250 students

in our sample we had 70 fresh.

Think about SE. (a measure of how much we would expect the # of fresh in our sample to vary as we take repeated samples)

$$SE = \sqrt{\#\text{draws}} \times SD_{box}$$

$$SD_{box} = (1 - 0)\sqrt{\text{fraction of tickets with 1} \times \text{fraction of tickets with 0}}$$

↑ unknown ↗

How to proceed?

- use the fractions in the sample.

$$SD_{box} = (1-0)\sqrt{\frac{70}{250} \times \frac{180}{250}} = 0.449$$

$$SE\ \#\ fresh. = \underbrace{\sqrt{250}}_{\sqrt{\text{sample size}}} \times \underbrace{0.449}_{SD_{box}} = 7.1$$

$$SE\ \%\ fresh = \frac{7.1}{250} \times 100 \approx 2.8\%.$$

(250 ← sample size)

Estimate of the population % $\approx$ sample %

$$= \frac{70}{250} \times 100 \approx 28\%.$$

So, from our sample of 250 students we estimate

% of freshpeople on campus is $28 \pm 2.8\%$.

- need to be a little bit careful in how we interpret this.

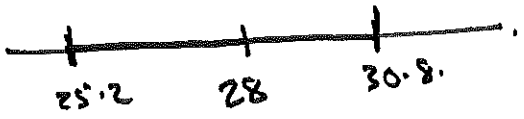
| Sample % | = | population % | + | chance error % |
|---|---|---|---|---|
| 28 | | ? | | ? |
| 28 | | 28 | | 0 |
| 28 | | 26 | | 2 |
| 28 | | 24 | | 4 |
| 28 | | 32 | | -4 |

What can we say about population % from a sample % of 28?

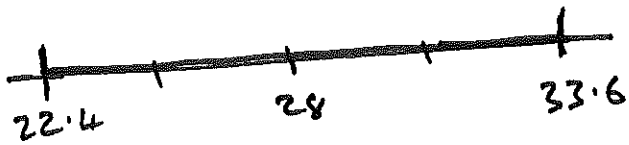
For each value of population % listed, we can think about the resulting chance error in terms of how many SE it is (ie standard units).

We can define a <u>confidence interval</u> around the sample % and use the normal approximation to define the <u>confidence level</u>

sample % ± 1 SE is 68% confidence interval




sample % ± 2 SE is 95% confidence interval




sample % ± 3 SE is 99.7 % confidence interval.

" we can be about 95% confident that the % of freshpersons on campus is between 22.4 and 33.6 % "

Note: we have not said that the chance of the % of fresh being between 22.4% and 33.6% is 0.95

because: probability is defined as the frequency of occurence of an event.

the population % is either between 22.4 and 33.6% or it is not.

this does not change however many times you measure it.

The chance is in the sampling, not in the parameter.

the parameter takes a fixed value.

$$\frac{4506}{15,125} \rightarrow 29.8\%.$$

From our sample, 70 out of 250 were fresh.

95% CI for population % of 22.4 – 33.6 %

In this case, the 95% CI covers the population %.

Consider a 2nd sample.

Size 250

# fresh 62

$$\text{sample } \% = \frac{62}{250} \times 100 = 24.8\%.$$

$$SD_{box} \approx (1-0)\sqrt{\frac{62}{250} \times \frac{188}{250}} = 0.43$$

↑ use sample proportion in place of unknown population proportion

SE on sum (# fresh) $= \sqrt{250} \times 0.43$

SE on % fresh $= \dfrac{\sqrt{250} \times 0.43}{250} \times 100 = 2.7\%$.

95% CI is $24.8\% \pm 2 \times 2.7\%$

or $19.4 \rightarrow 30.2\%$

again, 29.8 is in this range, so this <u>covers</u> the population %.

---

Now, repeat this 202 times...

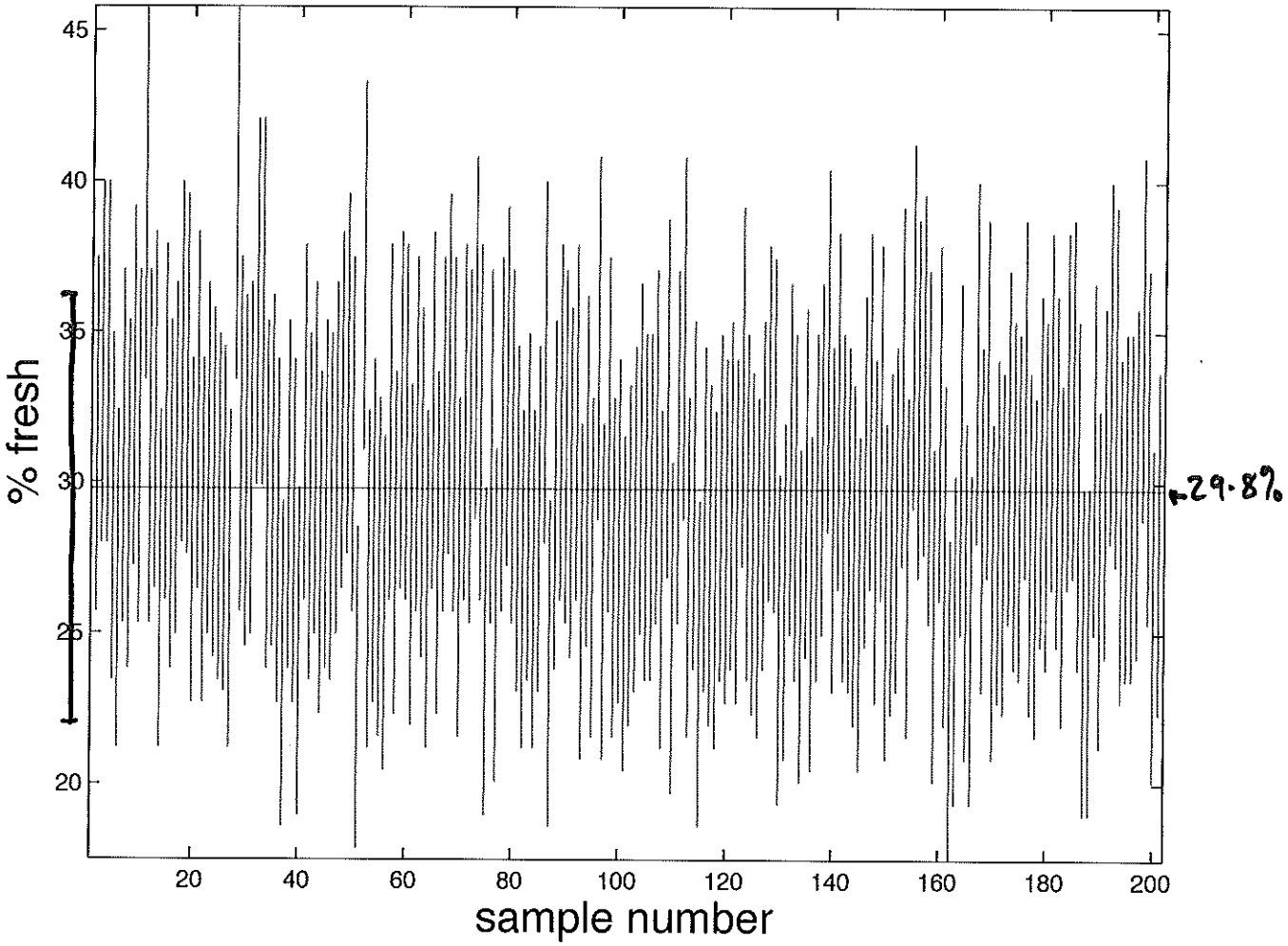
each sample will give a 95% CI.

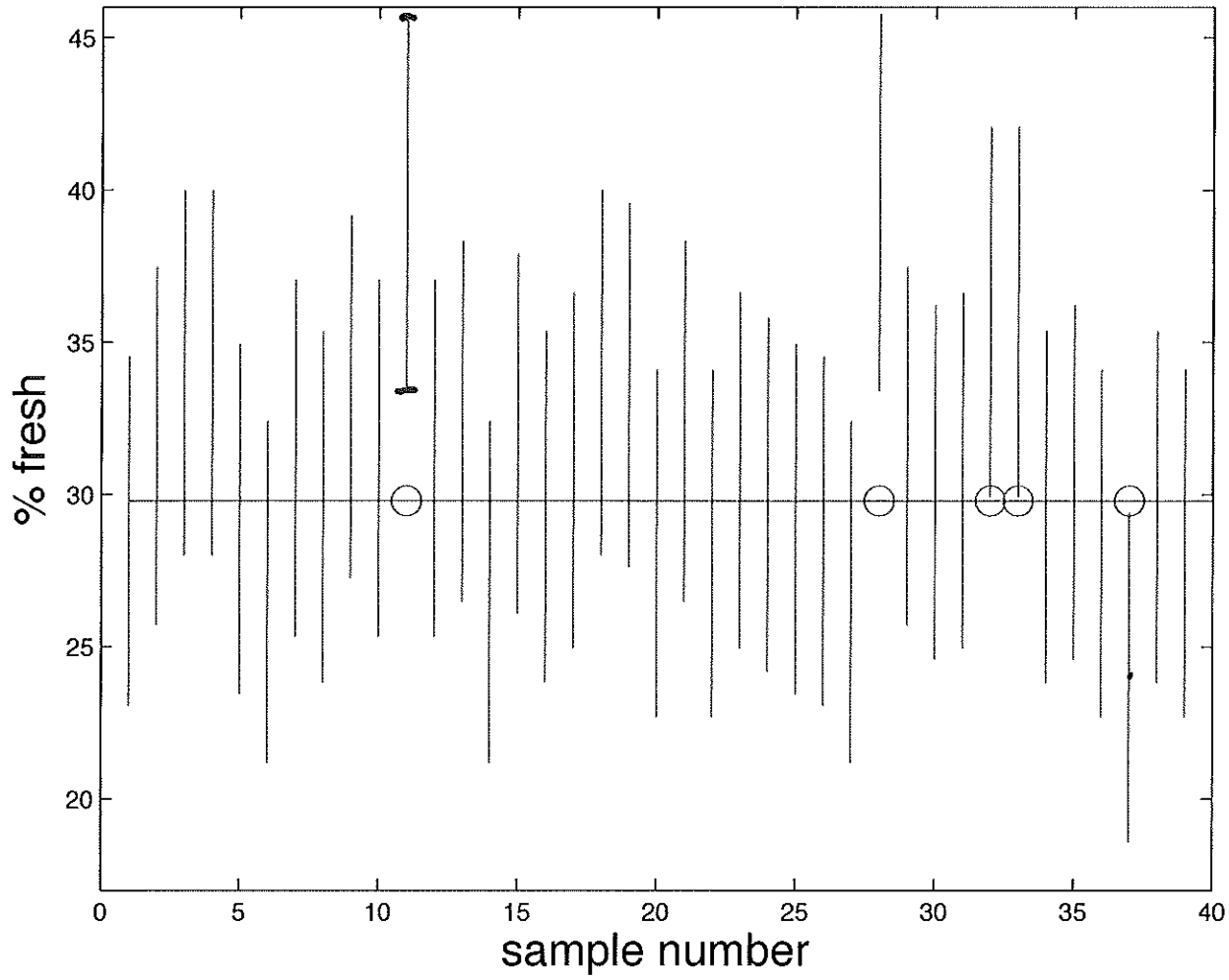
→ 95% of the CIs <u>cover</u> the true value.

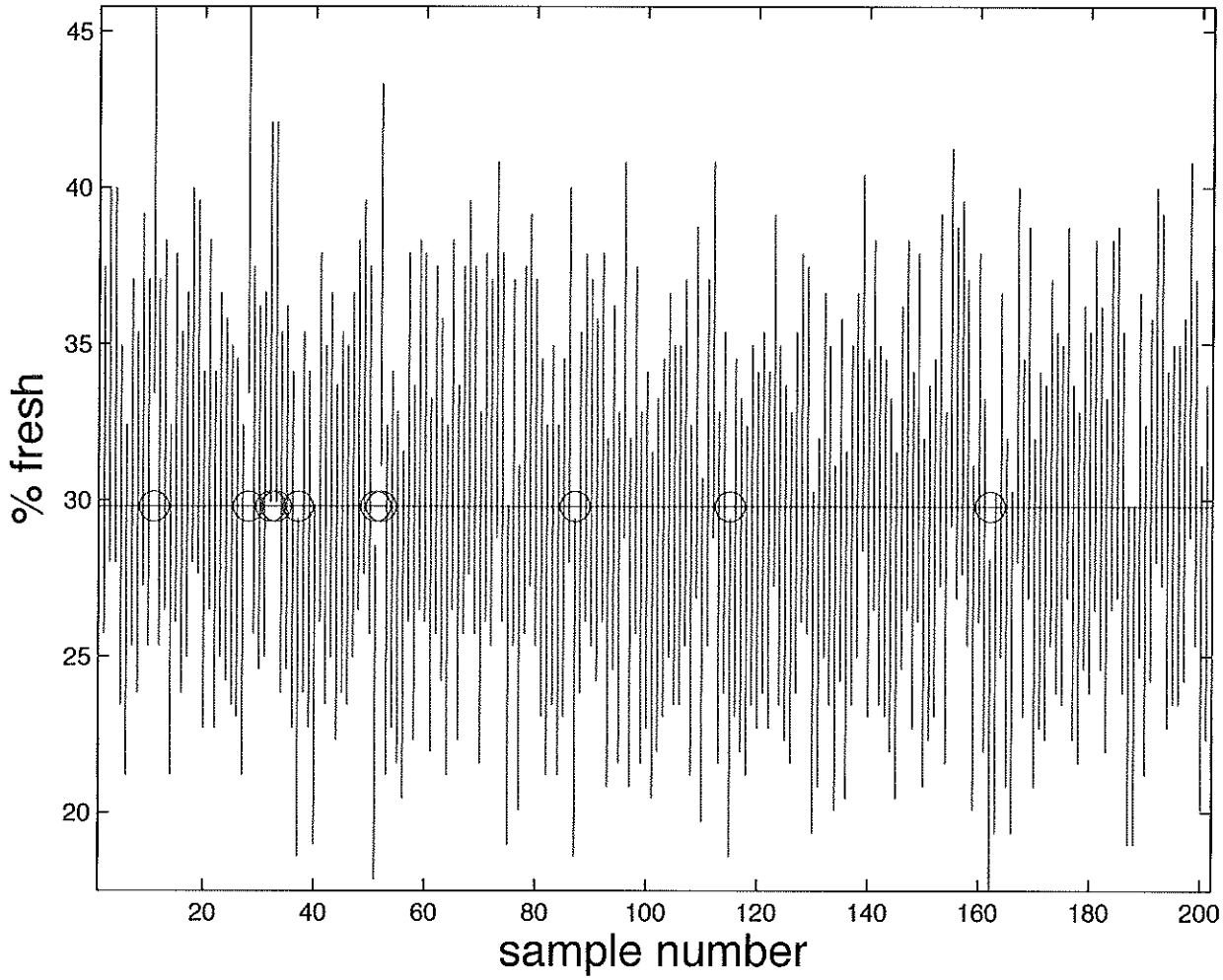
And <u>that</u> is what a 95% CI is:

95% of all samples will generate a CI that covers the true value

(<u>but</u>: with just one sample, you don't know if your sample does or does not).


45
40
35
30
25
20
% fresh
20
40
60
80
100
120
140
160
180
200
sample number
29.8%


% fresh
45
40
35
30
25
20
0
5
10
15
20
25
30
35
40
sample number


45
40
35
30
25
20
% fresh
20
40
60
80
100
120
140
160
180
200
sample number