Chance Errors in Sampling

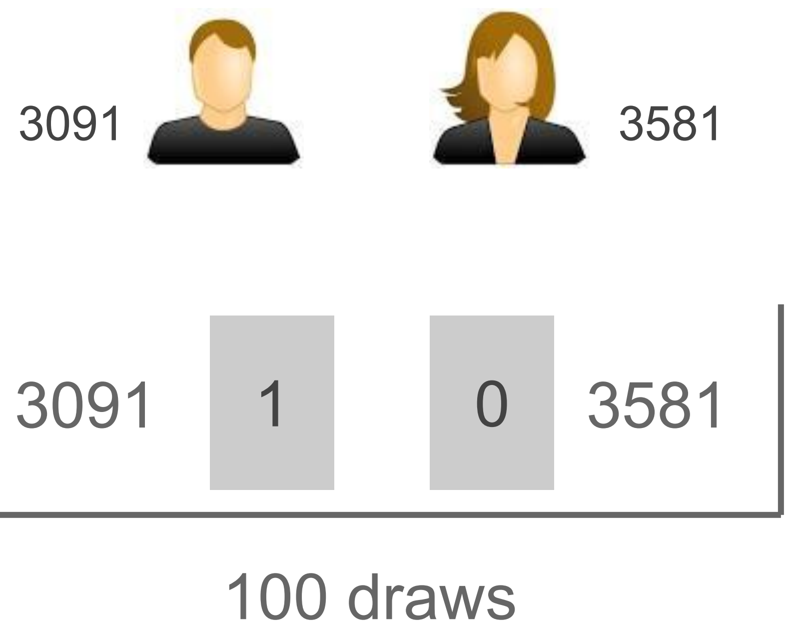
Gaston Sanchez

Sample surveys involve chance error.

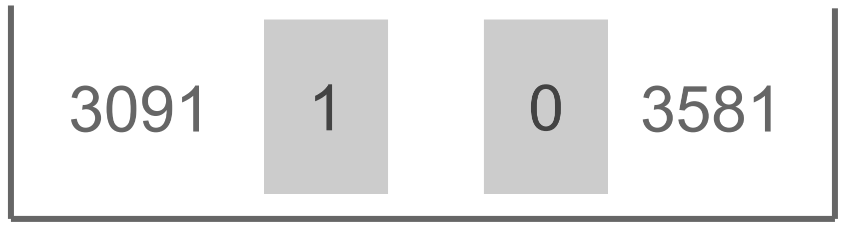
How to find the likely size of the chance error in a proportion for SRS?

Example: FPP page 359

Simple Random Sample of size 100 from a population of size 6672 individuals



Expected Value (for sum and percentage) of box



100 draws

# of draws = 100

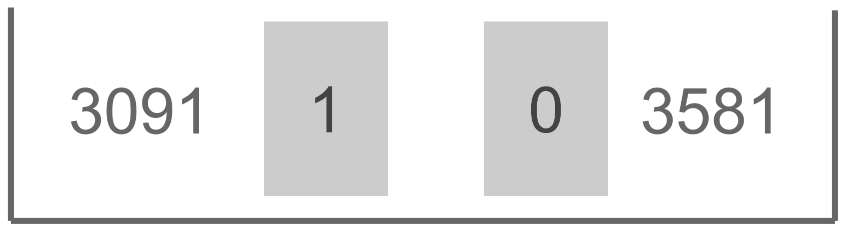
Avg of box = {3581(0) + 3091(1)} / 6672 = 0.4632

EV(sum) = (100) x (0.4632) = 46.32 men

EV(percent) = EV(sum) / 100 = 46.32% men

With a **SRS**, the expected value for the sample percentage equals the population percentage

Example page 359



100 draws

$$\text{SE for percentage} = \frac{\text{SE for sum}}{\text{Size of sample}} \times 100\%$$

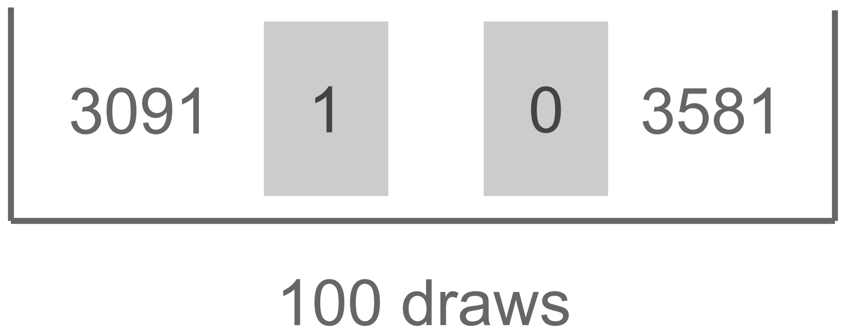
SE for number, and SD of box with 2 numbers

$$SE = \sqrt{\#\ \text{draws}}\ (\text{SD of box})$$

When a box has only two different numbers ("big" and "small"), the SD can be computed as:

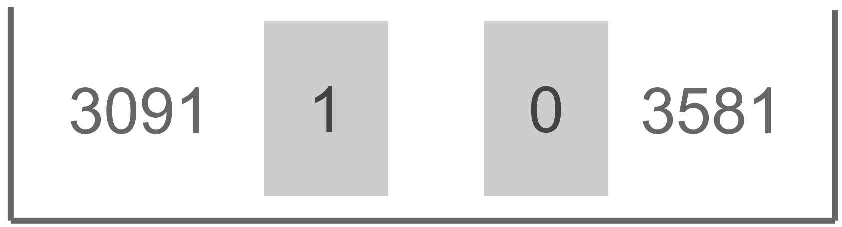
$$\left(\begin{array}{c}\text{big}\\ \text{number}\end{array} - \begin{array}{c}\text{small}\\ \text{number}\end{array}\right)\sqrt{\begin{array}{c}\text{fraction with}\\ \text{big number}\end{array} \times \begin{array}{c}\text{fraction with}\\ \text{small number}\end{array}}$$

Example page 359



$$\text{SE for percentage} = \frac{\text{SE for sum}}{\text{Size of sample}} \times 100\%$$

SE for percentage



100 draws

$$SD(box) = (1 - 0)\sqrt{0.46 \times 0.54} = 0.50$$

$$SE(sum) = \sqrt{100}\ (0.50) = 5$$

$$SE(percentage) = (5 / 100) \times 100\% = 5\%$$

What happens when the sample gets bigger?

SE for percentage

3091 | 1 | 0 | 3581

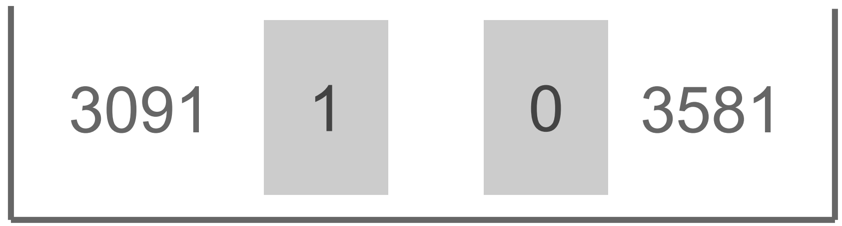
400 draws

$$SD(box) = (1 - 0)\sqrt{0.46 \times 0.54} = 0.50$$

$$SE(sum) = \sqrt{400}\ (0.50) = 10$$

$$SE(percentage) = (10 / 400) \times 100\% = 2.5\%$$

Example page 359



400 draws

SE for percentage = (5 / 100) x (100%) = 5%

Sampling men demo

More Normal Approximations

EV and SE for a sample percentage and how to use the normal curve to compute chances

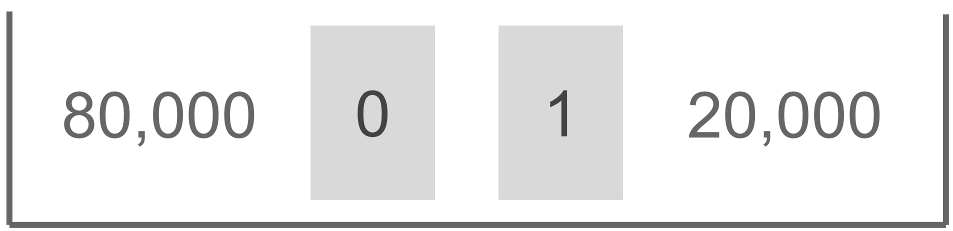
Example page 362-363

SRS of 400 from census data with

80,000
under $50k

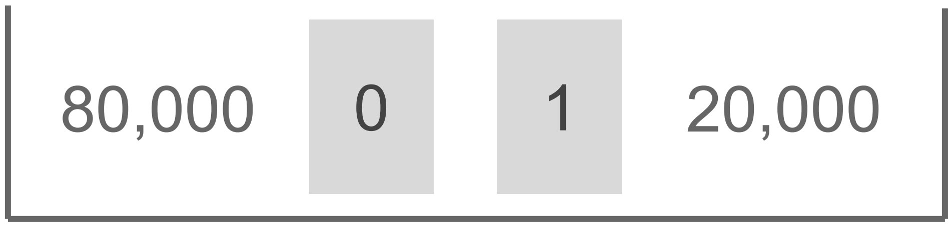
20,000
over $50k

What's the estimate percentage of "over $50k"?



400 draws

Example page 362-363



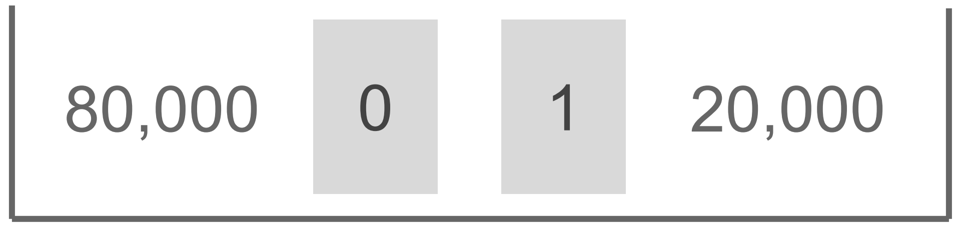
400 draws

# of draws = 400

Average of box = (20,000) / 100,000 = **0.2**

EV sum = (400) x (0.2) = **80**

Example page 362-363



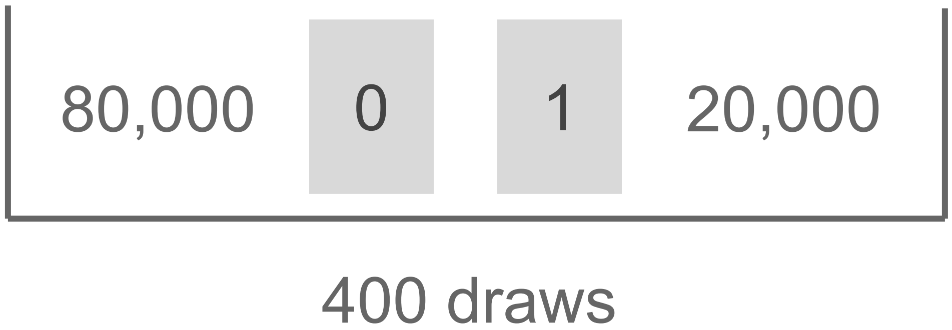
400 draws

# of draws = 400

SD box = $(1 - 0)\sqrt{(2/10) \times (8/10)} = 0.4$

SE sum = $\sqrt{400} \times (0.4) = \mathbf{8}$

Example page 362-363



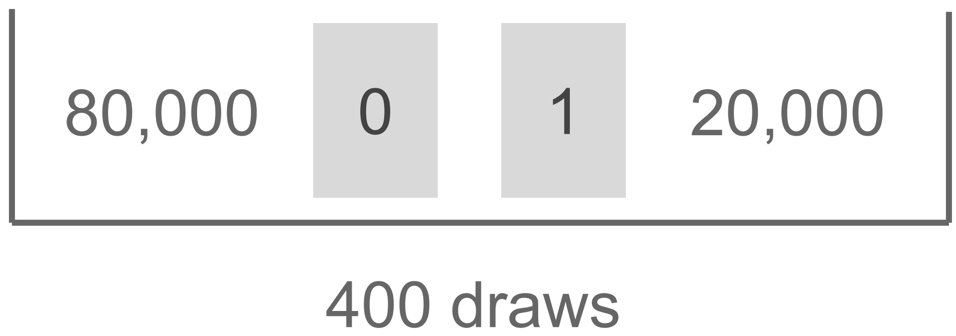
EV sum = **80** SE sum = **8**

EV % = (80 / 400) x 100% = 20%

SE % = (8 / 400) x 100% = 2%

When drawing at random from a box of 0’s and 1’s, the percentage of 1’s among the draws is likely to be around EV of %, give or take SE of % or so.

Example page 364



Estimate the chance that between 18% and 22% of the persons in the sample earn more then $50,000 a year

$$SU_1 = \frac{18 - 20}{2} = -1$$

$$SU_2 = \frac{22 - 20}{2} = 1$$

$$P(-1 < Z < 1) = 68\%$$

Correction Factor

When estimating percentages, it is the absolute size of the sample which determines accuracy

What's more accurate?

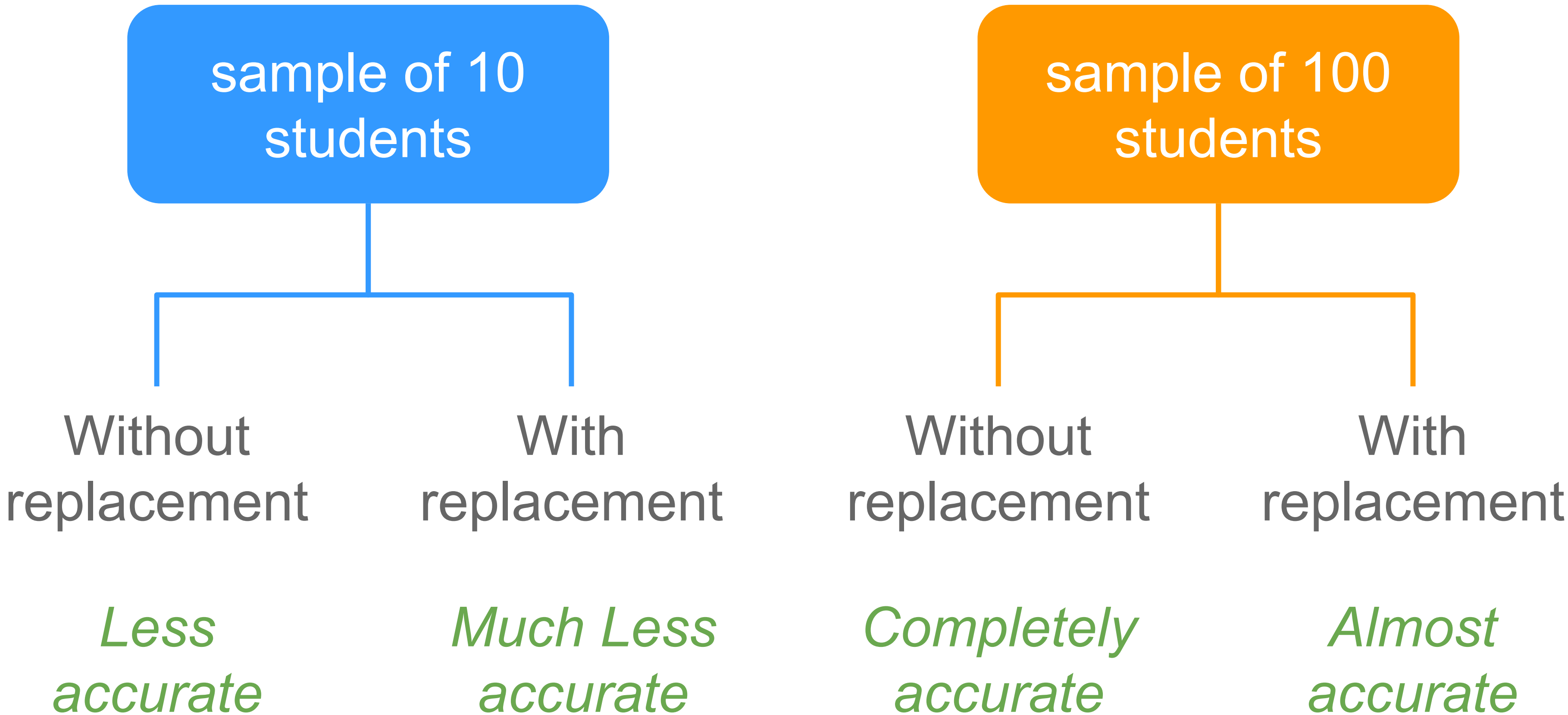
Statistic = Parameter + Chance Error + Bias

Sampling with replacement?

Sampling without replacement?

Small chance error

Number of Stat 2 students born in California



When drawing without replacement, the variability in the box is slightly reduced.

Finite Population Correction Factor

$$\text{SE w/o replacement} = \text{SE with replacement} \times \text{Correction factor}$$

$$\text{Correction factor} = \sqrt{\frac{\text{pop size} - \text{sample size}}{\text{pop size} - 1}}$$

Gaston Sanchez

Another way to express the Correction Factor

$$\sqrt{\frac{\text{Number of tickets in box - number of draws}}{\text{Number of tickets in box - one}}}$$

400 tickets in a box; EV and SE of sum?

300 0 1 *100*

100 draws, w/o replacement

Avg of box = (300x0 + 100x1) / 400 = 1 / 4

SD of box = $(1-0)\sqrt{(300/400)(100/400)} \approx 0.433$

EV = 100 (1/4) = 25

SE = $\sqrt{100}$ (0.433) = 4.33 (with replacement)

SE = $(4.33)\sqrt{(400 - 100) / (400 - 1)}$ = 3.75 (w/o replacement)

Special Cases for Correction Factor

Sample size = Population size
then **C.F.** = 0

Sample size = 1
then **C.F.** = 1

0 ≤ **C.F.** ≤ 1

Correction Factor with number of draws = 2500

Number of tickets in the box	Correction factor (to 5 decimal digits)
5,000	0.70718
10,000	0.86607
100,000	0.98743
500,000	0.99750
1,500,000	0.99917
15,000,000	0.99992

About Correction Factor

Correction Factor can be used whenever draws are made without replacement, for any **SE** (sum's, %'s, averages)

Use **C.F.** whenever sample size is at least 10% of population size

Less than 10% pop size, implies that the chance of asking the same person twice is small