Sampling of size n from population

- model: slips drawn from bag
- drawing

with replacement (i.i.d.)
without replacement (SRS)

$\Big\}$ very little difference if $\underbrace{n \ll \text{size of population}}$

if $\dfrac{n}{\text{size of population}} \le 0.1$

Think of independence as events not influencing one another.

Extend idea of independence to variables X, Y:

Say they are independent if they are not associated

Several facts about quantitative vars.

1. If X, Y are quant. vars and X has a mean μ_X, Y has a mean μ_Y, then their sum $X+Y$ has mean $\mu_X + \mu_Y$.

Their difference $X-Y$ has mean $\mu_X - \mu_Y$.

Ex.) If you're a golfer and average 91 strokes on course A and 83 strokes on course B. If you play both courses on one day, and take

X = your score on course A

Y = " " " " B

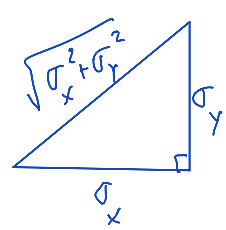
then

average sum $X+Y$ will be $\mu_X + \mu_X = 91 + 83 = 174$

" difference $X-Y$ " " $\mu_X - \mu_Y = 91 - 83 = 8$.

2. If X, Y are independent and X has a s.d. σ_X, Y has a standard deviation σ_Y, then both

$\left.\begin{matrix} X+Y \\ X-Y \end{matrix}\right\}$ have standard deviation $\sqrt{\sigma_X^2 + \sigma_Y^2}$.



Ex.)

Say Laura is a bowler whose scores have

mean $\mu = 161$

s.d. $\sigma = 14$

Laura bowls two games adding her scores

X_1 = score in 1^{st}

X_2 = " " 2^{nd}

Know

$S = X_1 + X_2$ has mean $161 + 161 = 322$

" s.d. $\sqrt{14^2 + 14^2} = \sqrt{2(14)^2} = 14\sqrt{2}$

$= \sigma\sqrt{2}$.

If she bowls 3 games, then her summed / total score

$X_1 + X_2 + X_3$ has mean $= 3\mu = 3(161)$

s.d. $= \sqrt{14^2 + 14^2 + 14^2} = 14\sqrt{3}$.

3, Corollary to 2:

If $X_1, X_2, \ldots, X_n$ is an i.i.d. sample from a population with mean μ, s.d. σ, then the

a) mean for the sum: $n\mu$, s.d. for sum: $\sigma\sqrt{n}$

b) mean for $\frac{1}{n}(X_1 + X_2 + \cdots + X_n)$ (i.e., the mean for the sample mean $\bar{X}$)

$\mu_{\bar{X}} = \mu$ (sample means are unbiased estimators of the population mean)

and the s.d. of $\bar{X}$

$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$.

Ex.

If Laura bowls 3 games, expect her average of the three

$$\bar{X} = \frac{1}{3}(X_1 + X_2 + X_3)$$

to have a distribution with mean 161 and s.d. $\frac{14}{\sqrt{3}}$

If she bowls 20 games, then her average

$$\bar{X} = \frac{1}{20}(X_1 + X_2 + \cdots + X_{20})$$

will have mean = 161 and s.d. $\frac{14}{\sqrt{20}}$.

Central Limit Theorem:

1) The sum of $X_1, \ldots, X_n$ (i.i.d. sample from quantitative population) has an approximate normal distribution for large enough n (and given what we learned above, that normal dist. will have mean $n\mu_X$ and s.d. $\sigma_X\sqrt{n}$).

2) The average $\frac{1}{n}(X_1 + \cdots + X_n)$ will likewise be approximately normal for large enough n (with mean μ_X, s.d. = $\frac{\sigma_X}{\sqrt{n}}$).

Both claims assume i.i.d. samples $X_1, \ldots, X_n$ taken from a population. They also apply for SRS's, if rule of thumb (sampling less than 10% of population).

```
Stat 145, Mon 18-Oct-2021 -- Mon 18-Oct-2021
Biostatistics
Spring 2021


-------------------------------
Monday, October 18th 2021
-------------------------------
Wk 8, Mo
Topic:: Central Limit Theorem
Read::  Lock5 5.2

Variables can
 - have an association, or
 - not have an association.
    We also talk about independent variables, which is roughly
    the same as


   Examples:
    1. If we draw twice from a bag and take
         X = 1st outcome
         Y = 2nd outcome
       then X and Y are
        i) independent if sampling "with replacement"
           call this an i.i.d. random sample of size 2

       ii) approximately independent if the composition of the bag is
           little changed after the first draw

    2. If we draw n times from a bag and take
         X_1 = 1st outcome
         X_2 = 2nd outcome
             .
             .
             .
         X_n = nth outcome
       the X_i are
        i) independent if sampling "with replacement"
           call this an i.i.d. random sample of size n
```

```
ii) approximately independent if the composition of the bag is
    little changed after by the draws
```

Rule of thumb: Size n of sample is ≤ 10% of size of bag's contents

```
A random variable X is one that is numeric for each case
 - sex: F/M we think of as categorical
 - X(case) = 0 if case=female, 1 if case=male is a random variable
```

(binary)

(turns outcomes into numbers)

```
Some facts about independent normal random variables
 - If X and Y are independent normal random variables, with
    X ~ Norm(mu_1, sigma_1)
    Y ~ Norm(mu_2, sigma_2)
   then X+Y (their sum) is ~ Norm(mu_1 + mu_2, sqrt(sigma_1^2 + sigma_2^2)
   then X+Y (their difference) is ~ Norm(mu_1-mu_2, sqrt(sigma_1^2 + sigma_2^2))
```

— means "has a normal dist. with mean μ_1, sd σ_1"

X-Y

```
   Ex.: Suppose Ray and Joan are bowlers.  Their scores have normal dists
         R ~ Norm(142, 17)
         J ~ Norm(138, 22)
        How likely is it for them, in one game, to have a combined score > 350?
```

R + J ~ Norm(280, 27.803)

Answer comes from 1 − pnorm(350, 280, 27.803)

```
 - If we draw an i.i.d. random sample of size n, each X_i ~ Norm(mu, sigma),
   then the
     sum = X_1 + ... + X_n  is  Norm(n mu, sigma sqrt(n))
     avg = (X_1 + ... + X_n) / n  is  Norm(mu, sigma / sqrt(n))


Central Limit Theorem
 Suppose a random sample of size n is drawn from the population either
  - with replacement (so it is i.i.d.), or
  - with n smaller than 10% of the full population.
 If the variable of interest is quantitative and n is large enough, then
   the sum X_1 + ... + X_n is approximately normal
   the mean (X_1 + ... + X_n)/n is approximately normal
 If the variable of interest is binary categorical and n is large enough,
 then the sample proportion has approximately a normal distribution.
```

```
Explorations using apps at
 https://onlinestatbook.com/stat_sim/sampling_dist/index.html
 https://shiny.calvin.edu:3838/scofield/samplingDists/
 https://shiny.calvin.edu:3838/scofield/cltProportions/
```

Central Limit Theorem

In summary, here is the take-away from the **Central Limit Theorem**.

Suppose you have a random sample of size n that is either

- i.i.d., or
- an SRS, with the sample size n being no more than 10% of the size of the population.

In the case that

1. the variable under consideration is quantitative, having population mean μ and standard deviation σ, then the sampling distribution for the sample mean $\overline{X}$ is approximately $\mathsf{Norm}(\mu, \sigma/\sqrt{n})$ for n large enough.

2. the variable under consideration is binary categorical, having population proportion p, then the sampling distribution for the sample proportion $\widehat{p}$ is approximately $\mathsf{Norm}(p, \sqrt{p(1-p)/n})$ for n large enough.

Since

- null distributions
- randomization distributions
- bootstrap distributions

are all specialized versions of sampling distributions, then so long as the sample statistic in question is the sample's *mean* $\overline{X}$ or the sample *proportion* $\widehat{p}$, we can expect the CLT to apply to these as well.