008 - Central Limit Theorem

EPIB 607 - FALL 2020

Sahir Rai Bhatnagar
Department of Epidemiology, Biostatistics, and Occupational Health
McGill University

sahir.bhatnagar@mcgill.ca

slides compiled on September 23, 2020

Statistical Concepts and Prinicples

Central Limit Theorem

Standard deviation and variance of a random variable Y

- $Y \sim \text{unknown\_distribution}(\mu, \sigma)$
- Standard Deviation σ, and Variance σ^2, of a random variable Y with mean μ.

$$\begin{aligned}
Var[Y] &= \sigma^2 = \text{mean of } (Y - \mu)^2 \\
SD[Y] &= \sigma \\
Var[Y \pm a\ constant] &= Var[Y] \\
SD[Y \pm a\ constant] &= \sigma \\
Var[Y \times a\ constant] &= constant^2 \ \times \ Var[Y] \\
SD[Y \times a\ constant] &= |constant| \times \sigma
\end{aligned}$$

Rules for Variances and SDs of sums and means of *n* independent random variables

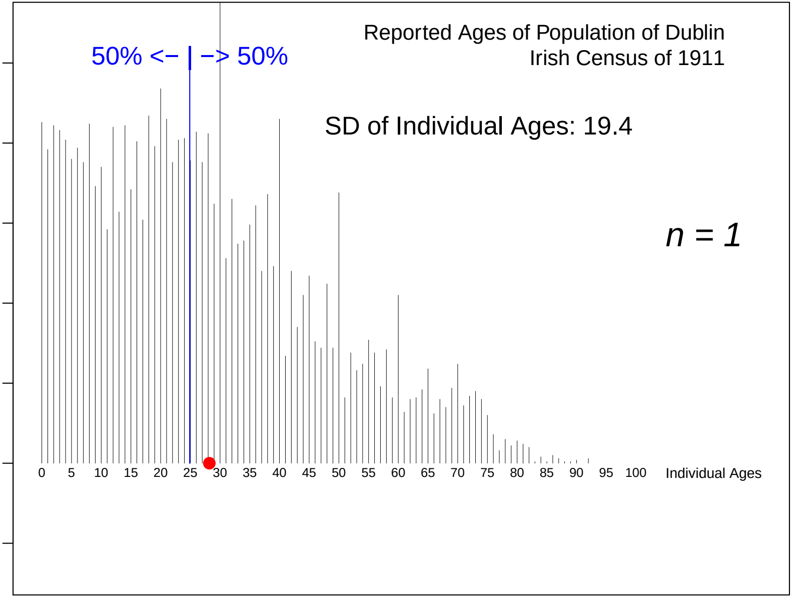
Sums

$$Var[Y_1 + Y_2 + \cdots + Y_n] = \sigma^2 + \sigma^2 + \cdots + \sigma^2 = n \times \sigma^2$$
$$SD[Y_1 + Y_2 + \cdots + Y_n] = \sqrt{n} \times \sigma$$

Means

$$Var\left[\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right] = \frac{1}{n} \times \sigma^2$$
$$SD\left[\frac{Y_1 + Y_2 + \cdots + Y_n}{n}\right] = \sqrt{\frac{1}{n}} \times \sigma$$

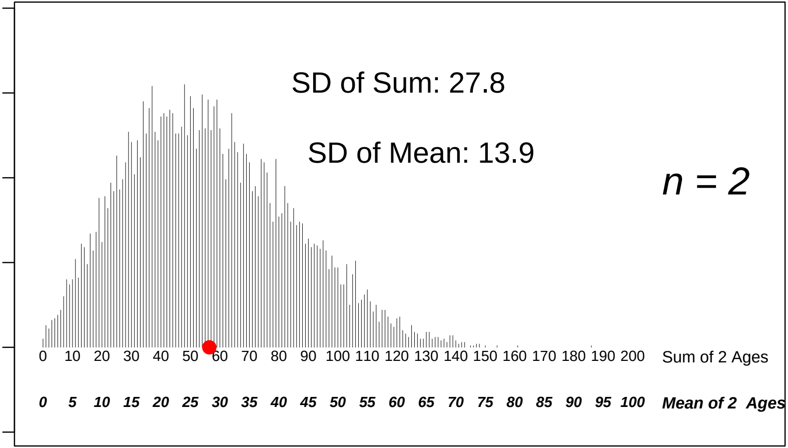
Age-distribution of the entire population of Dublin[1]



[1] http://www.census.nationalarchives.ie/help/about19011911census.html

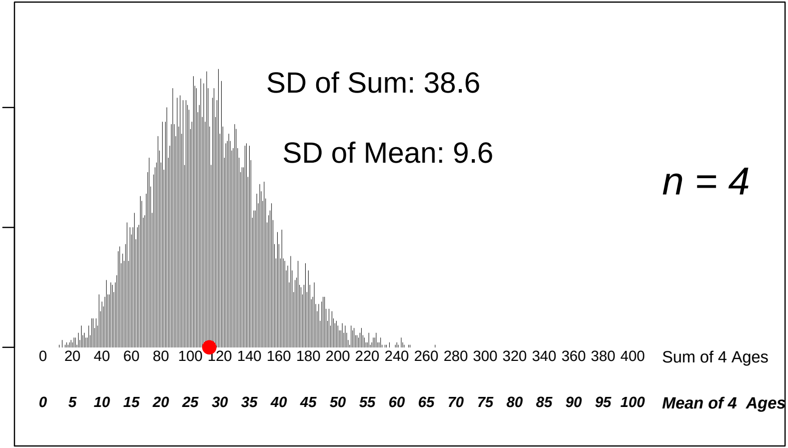
Distribution of 10000 Bootstrap samples of size 2

```
## [1] Ages of sampled persons in first 2 samples of size 2
##      [,1] [,2]
## [1,]   87   29
## [2,]    8   30
```



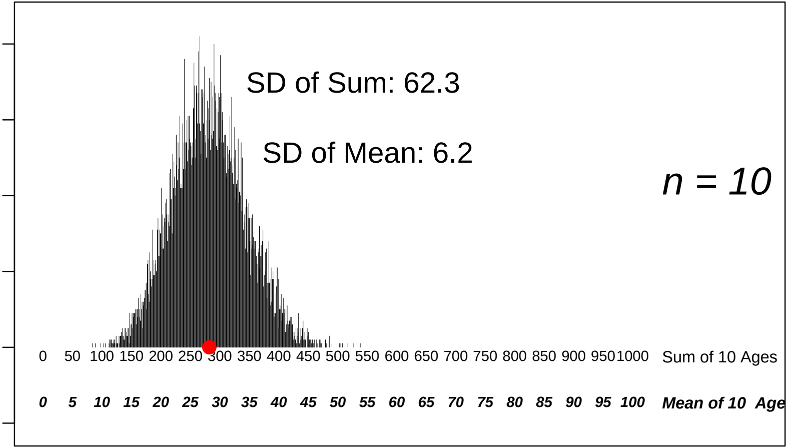
Distribution of 10000 Bootstrap samples of size 4

```
## [1] Ages of sampled persons in first 2 samples of size 4
##      [,1] [,2] [,3] [,4]
## [1,]   39   28    1   20
## [2,]   45    0   59   22
```

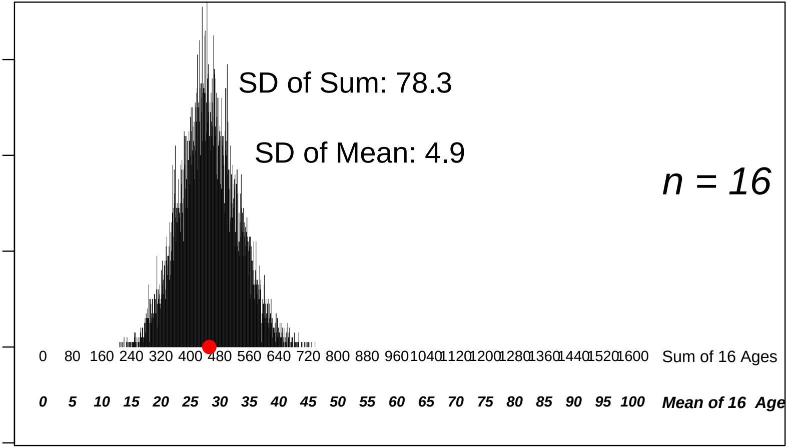


Distribution of 10000 Bootstrap samples of size 10

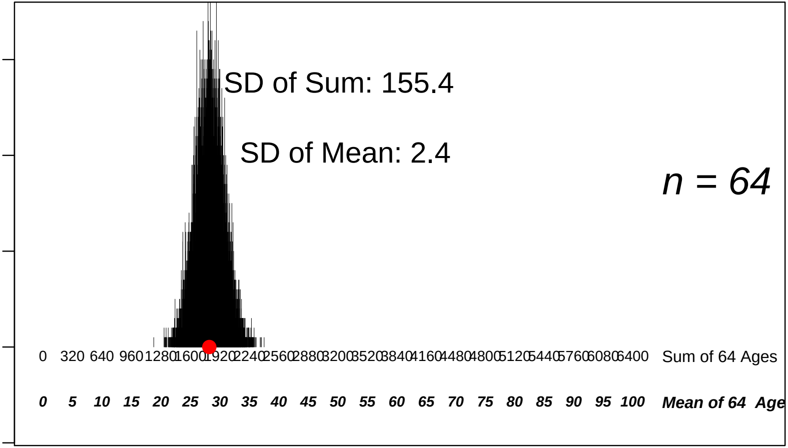
```
## [1] Ages of sampled persons in first 2 samples of size 10
##       [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]    48   22   12   42   46   28   35   40   27    33
## [2,]    14   18   37    0    5   13    3   12    4    45
```



Distribution of 10000 Bootstrap samples of size 16



Distribution of 10000 Bootstrap samples of size 64



Exercises

1. Based on the numbers in the 5 panels, derive the statistical law that connects the spreads of the sampling distributions of the sample sum to the spread of the individual ages. (Since the calculated sd's are based on a finite set of simulations, the numbers may not fit the law exactly ; also, the sd's shown are rounded)
2. Likewise, state the statistical law that connects the spreads of the sampling distributions of the sample mean to the spread of the individual ages. Use this law to predict the spread of the sampling distribution if we were to use a sample size of $n = 100$.
3. What n would you need to have so that the (approx. 95%) Margin of Error, i.e., 2 times the SD of the mean (or 2 times the 'Standard Error of the Mean' or 2 times the 'SEM') is less than (a) 1 year (b) 0.5 years?

Statistical Concepts and Prinicples

Central Limit Theorem

Properties of the sample mean: The Central Limit Theorem (CLT)

The sampling distribution of $\overline{Y}$ is Normal if Y is Normal. What probability distribution does the sample mean follow if Y is not Normal?

As sample size increases, the distribution of $\overline{Y}$ becomes closer to a Normal distribution, no matter what the distribution of sampled variable Y!

(This is true as long as the distribution has a finite variance.)

The Central Limit Theorem (CLT)

- The sampling distribution of $\bar{y}$ is, for a large enough n, close to Gaussian in shape no matter what the shape of the distribution of individual Y values.
- This phenomenon is referred to as the CENTRAL LIMIT THEOREM
- The CLT applied also to a sample proportion, slope, correlation, or any other statistic created by aggregation of individual observations

Theorem (Central Limit Theorem)

$$\textit{if } Y \sim ???(\mu_Y, \sigma_Y), \textit{ then}$$

$$\bar{y} \sim \mathcal{N}(\mu_Y, \sigma_Y/\sqrt{n})$$

Standard error (SE) of a sample statistic

- Recall: When we are talking about the variability of a **statistic**, we use the term **standard error** (not standard deviation). The standard error of the sample mean is $\sigma/\sqrt{n}$.

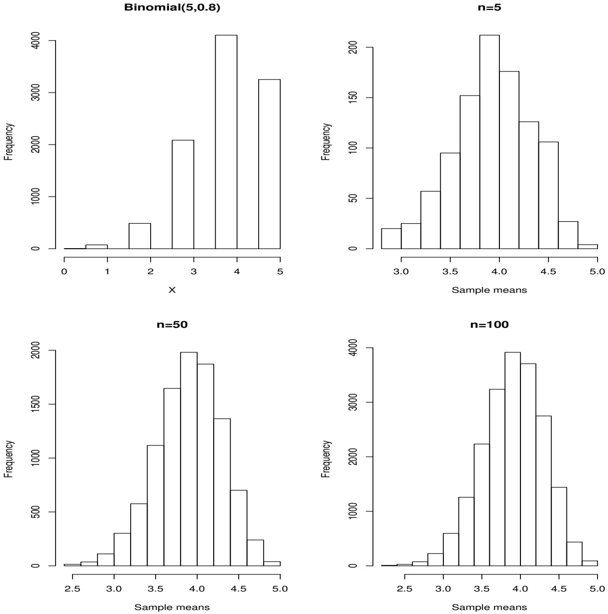
Remark (SE vs. SD)

In quantifying the instability of the sample mean ($\bar{y}$) statistic, we talk of SE of the mean (SEM)

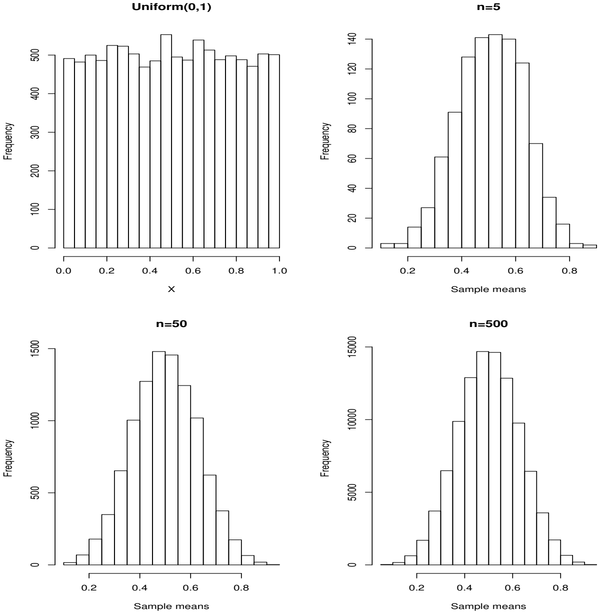
SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

SD(y) describes how far an individual y (typically) deviates from μ (or from $\bar{y}$).

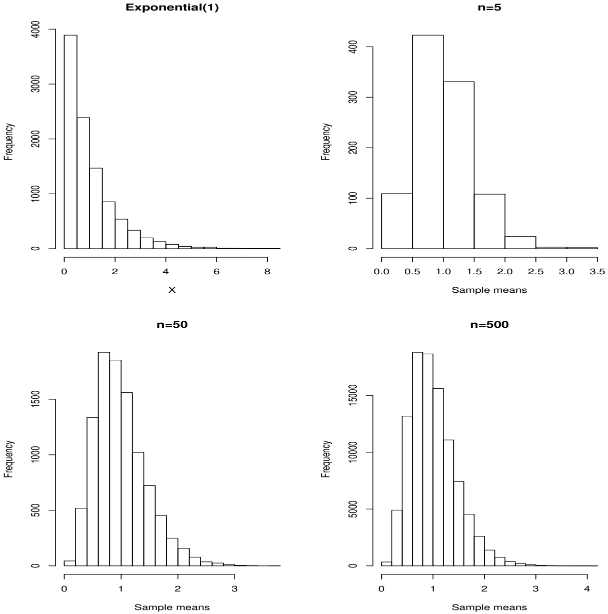
CLT in action: Binomial(n = 5,p = 0.8) distribution



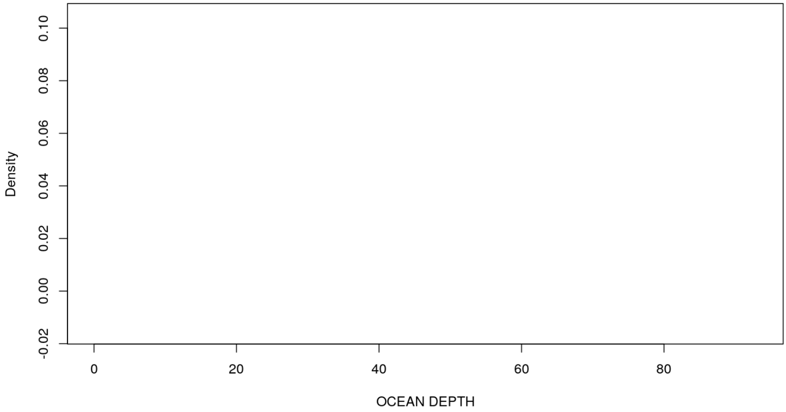
CLT in action: Uniform(a = 0, b = 1) distribution



CLT in action: Exponential($\lambda = 1$) distribution

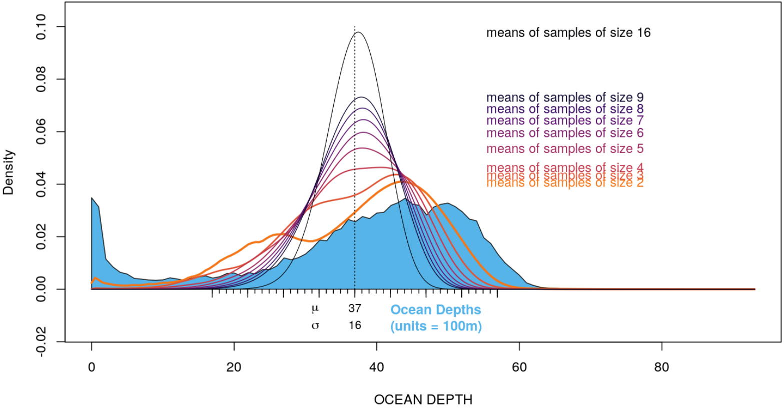


CLT in action: Depths of the ocean



How long does it take for the CLT to ’kick in’?

- How *fast* or slowly the CLT will kick in is a function of how symmetric, or how asymmetric and CLT-unfriendly, the distribution of Y (the depths of the ocean) is



Quadruple the work, half the benefit

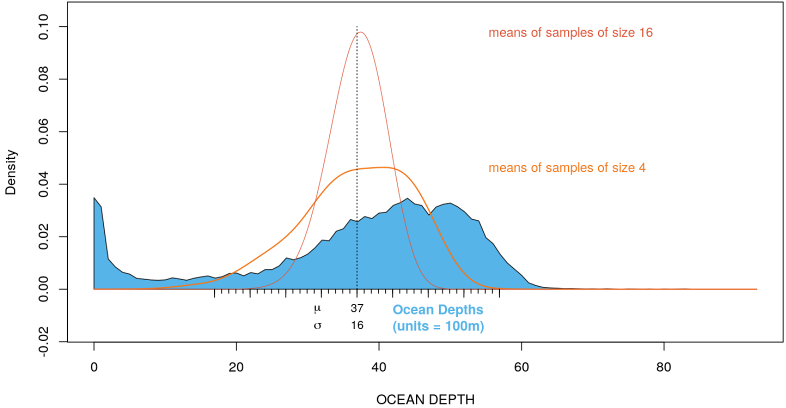


Figure: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

Session Info

```
R version 4.0.2 (2020-06-22)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 20.04 LTS

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/liblapack.so.3

attached base packages:
[1] tools     stats     graphics  grDevices utils     datasets  methods  
[8] base     

other attached packages:
 [1] NCStats_0.4.7   FSA_0.8.30      forcats_0.5.0   stringr_1.4.0  
 [5] dplyr_1.0.2     purrr_0.3.4     readr_1.3.1     tidyr_1.1.2    
 [9] tibble_3.0.3    ggplot2_3.3.2   tidyverse_1.3.0 knitr_1.29     

loaded via a namespace (and not attached):
 [1] ggdendro_0.1.22    httr_1.4.2         jsonlite_1.7.1     splines_4.0.2     
 [5] carData_3.0-4      modelr_0.1.8       assertthat_0.2.1   highr_0.8         
 [9] blob_1.2.1         ggstance_0.3.4     cellranger_1.1.0   mosaic_1.7.0      
[13] ggrepel_0.8.2      pillar_1.4.6       backports_1.1.9    lattice_0.20-41   
[17] glue_1.4.2         digest_0.6.25      polyclip_1.10-0    rvest_0.3.6       
[21] colorspace_1.4-1   htmltools_0.5.0    Matrix_1.2-18      plyr_1.8.6        
[25] pkgconfig_2.0.3    broom_0.7.0        haven_2.3.1        scales_1.1.1      
[29] tweenr_1.0.1       openxlsx_4.1.5     mosaicData_0.20.1  rio_0.5.16        
[33] TeachingDemos_2.12 ggforce_0.3.2      generics_0.0.2     farver_2.0.3      
[37] car_3.0-9          ellipsis_0.3.1     withr_2.2.0        cli_2.0.2         
[41] magrittr_1.5       crayon_1.3.4       readxl_1.3.1       evaluate_0.14     
[45] fs_1.5.0           fansi_0.4.1        MASS_7.3-53        xml2_1.3.2        
[49] foreign_0.8-79     data.table_1.13.0  hms_0.5.3          lifecycle_0.2.0   
[53] munsell_0.5.0      reprex_0.3.0       zip_2.1.1          compiler_4.0.2    
[57] rlang_0.4.7        grid_4.0.2         rstudioapi_0.11    htmlwidgets_1.5.1 
[61] crosstalk_1.1.0.1  mosaicCore_0.8.0   gtable_0.3.0       abind_1.4-5       
[65] DBI_1.1.0          curl_4.3           R6_2.4.1           gridExtra_2.3     
[69] lubridate_1.7.9    ggformula_0.9.4    stringi_1.5.3      Rcpp_1.0.5        
[73] vctrs_0.3.4        leaflet_2.0.3      dbplyr_1.4.4       tidyselect_1.1.0  
[77] xfun_0.17         
```