012 - Central Limit Theorem

EPIB 607

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

sahir.bhatnagar@mcgill.ca

slides compiled on September 26, 2021

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

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

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

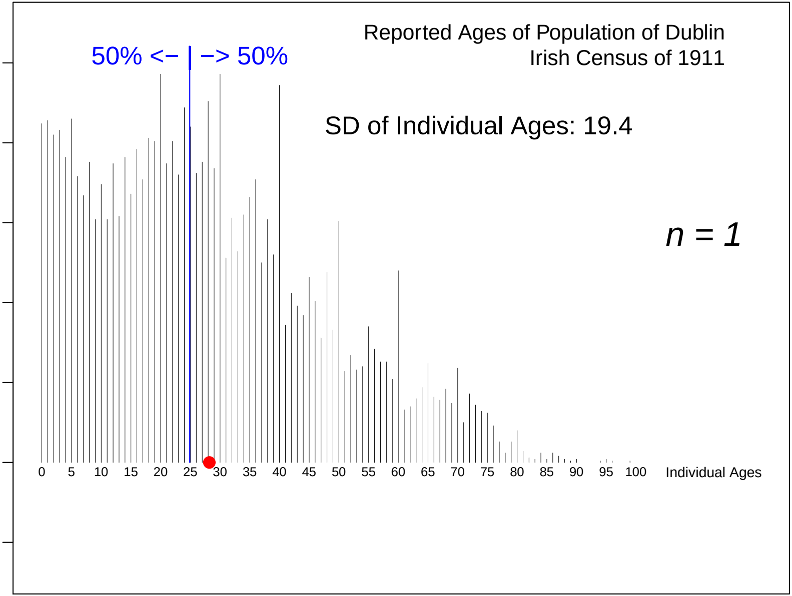
Sums

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

Means

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

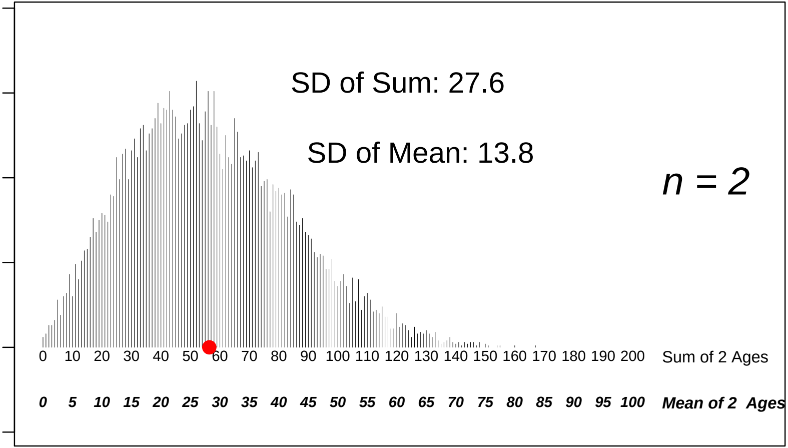
Age-distribution of the entire population of Dublin[1]



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

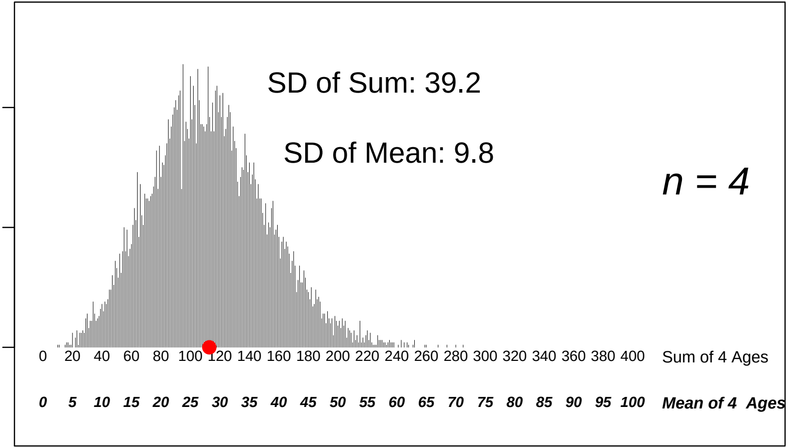
Distribution of 10000 Bootstrap samples of size 2

```
## [1] Ages of sampled persons in first 2 samples of size 2
##      [,1] [,2]
## [1,]   24   11
## [2,]   65    4
```



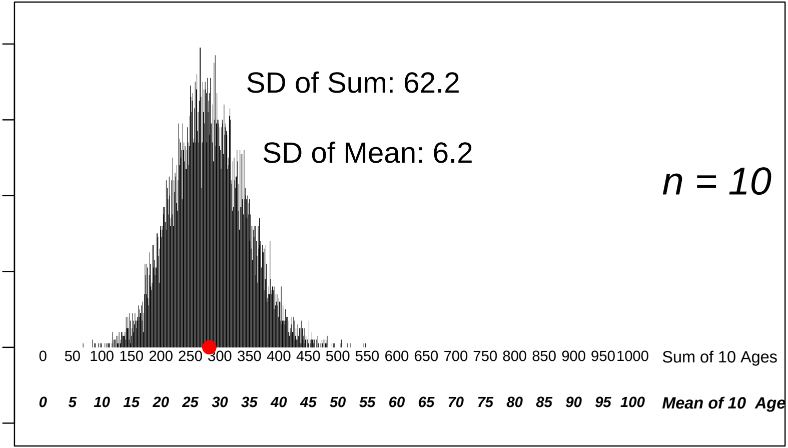
Distribution of 10000 Bootstrap samples of size 4

```
## [1] Ages of sampled persons in first 2 samples of size 4
##      [,1] [,2] [,3] [,4]
## [1,]   60   53   14   50
## [2,]   58   23    2   35
```

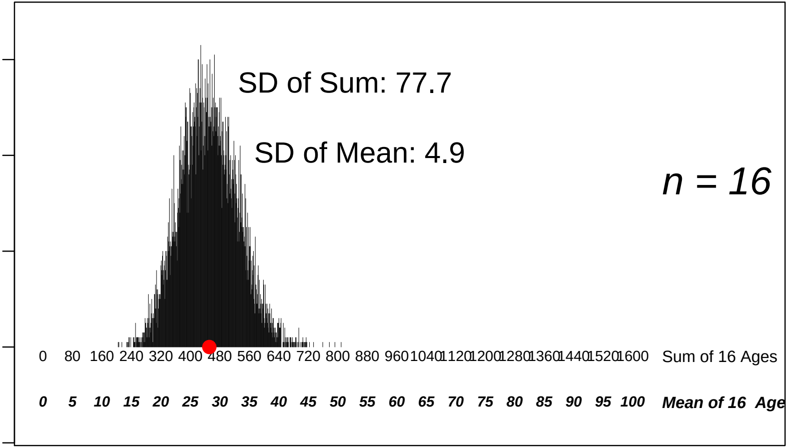


Distribution of 10000 Bootstrap samples of size 10

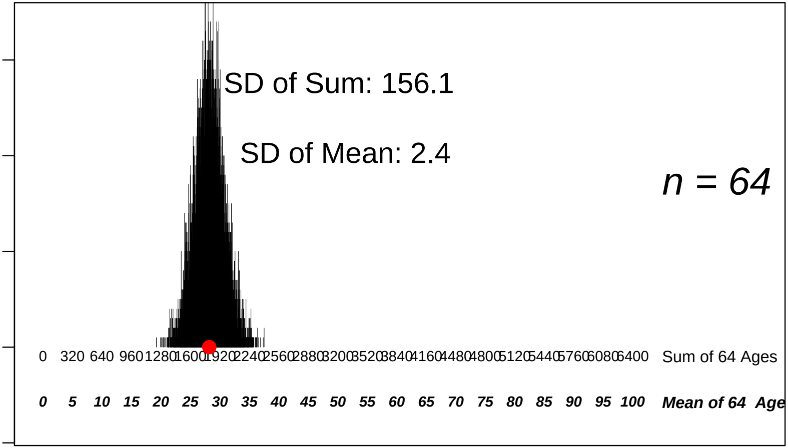
```
## [1] Ages of sampled persons in first 2 samples of size 10
##      [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
## [1,]   42   11   16   18    3   48   28    1    7     5
## [2,]   23   20   20   29   38   15   82   68   18    16
```



Distribution of 10000 Bootstrap samples of size 16



Distribution of 10000 Bootstrap samples of size 64



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?

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 1 (Central Limit Theorem).

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

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

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 $\sigma/\sqrt{n}$.

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 $\sigma/\sqrt{n}$.

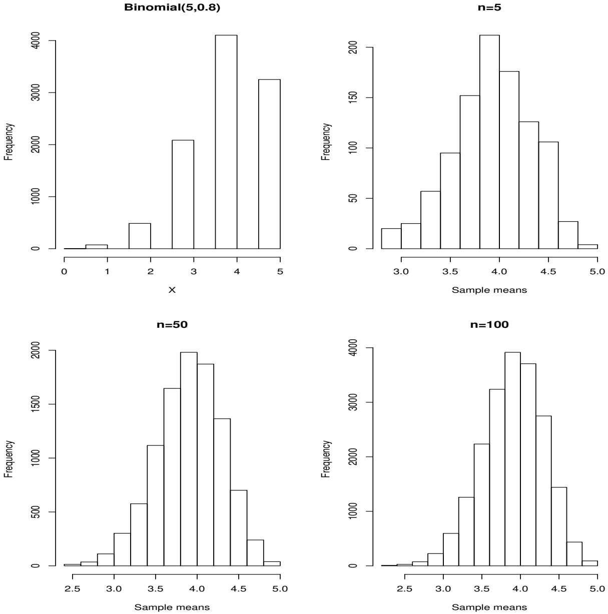
Remark 1 (SE vs. SD).

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

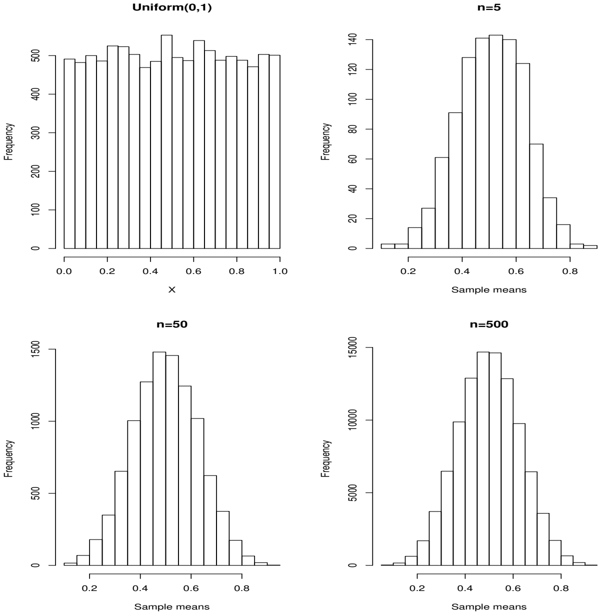
SE($\bar{y}$) describes how far $\bar{y}$ could (typically) deviate from μ;

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

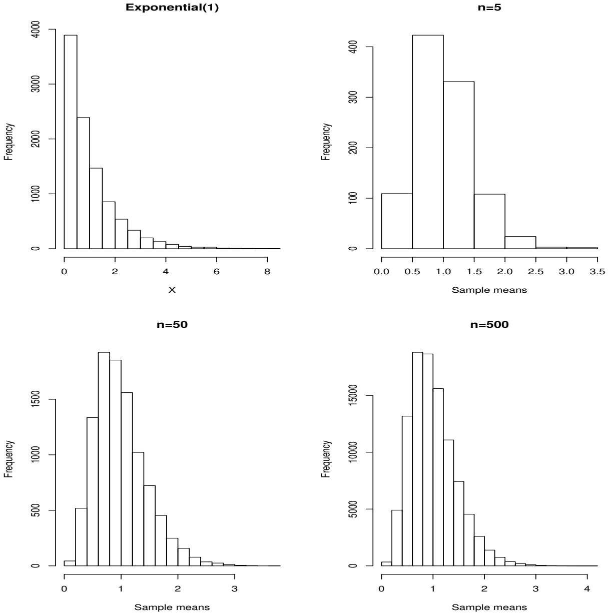
CLT in action: Binomial(n = 5,p = 0.8) distribution



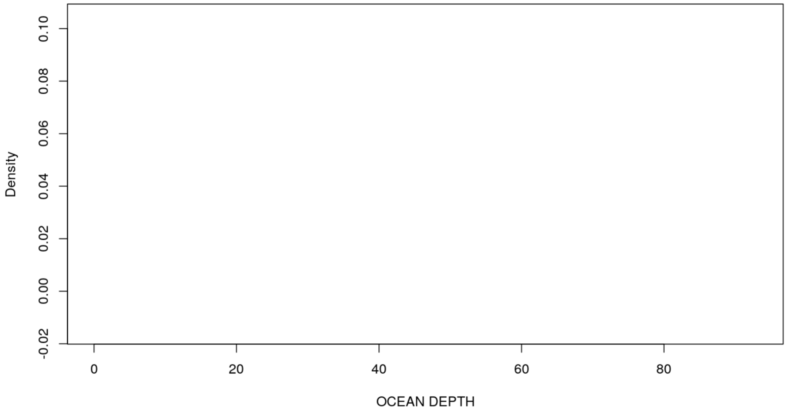
CLT in action: Uniform(a = 0, b = 1) distribution



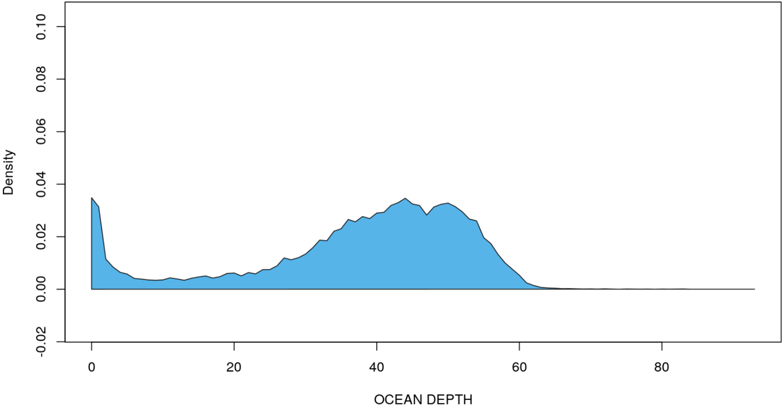
CLT in action: Exponential($\lambda = 1$) distribution



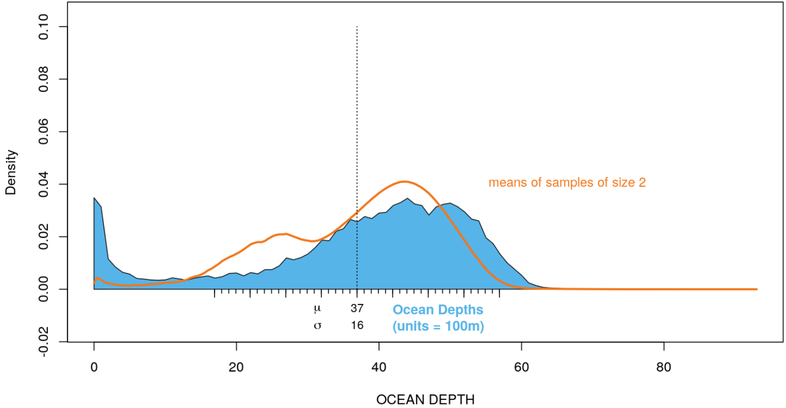
CLT in action: Depths of the ocean



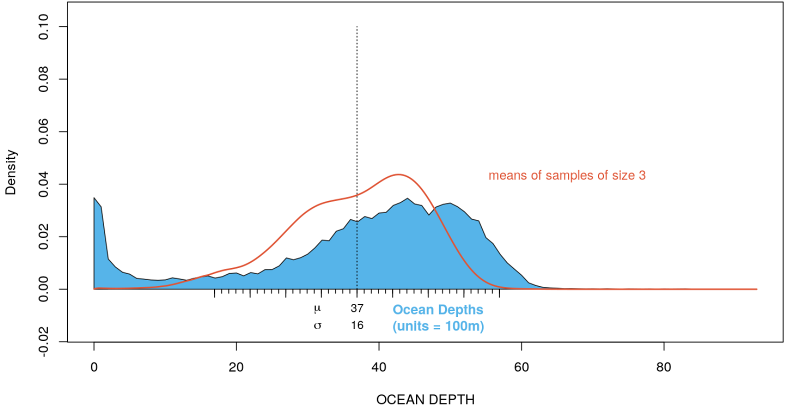
CLT in action: Depths of the ocean



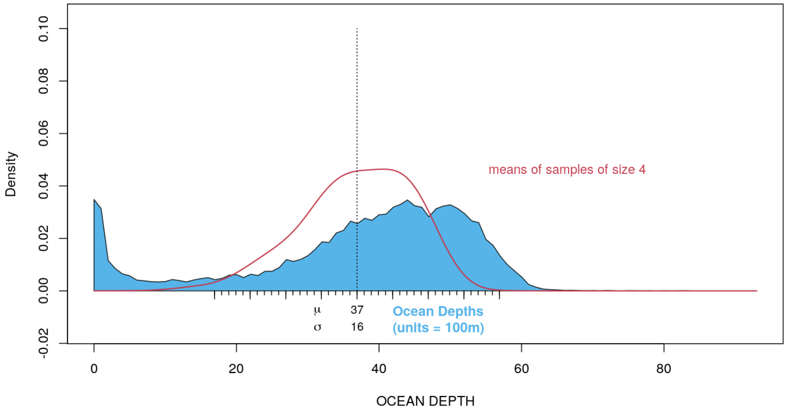
CLT in action: Depths of the ocean



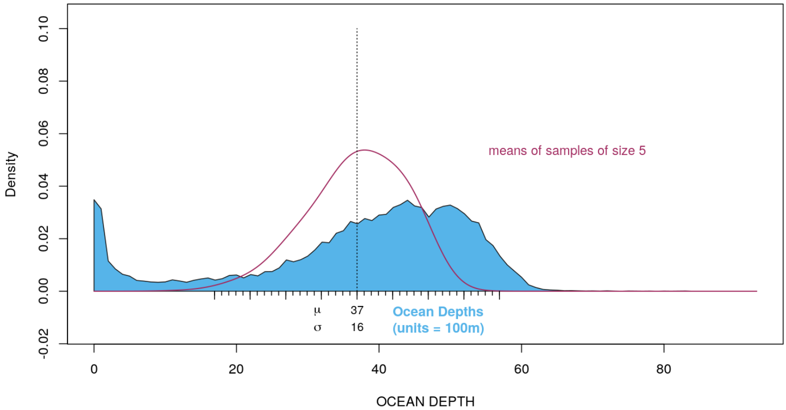
CLT in action: Depths of the ocean



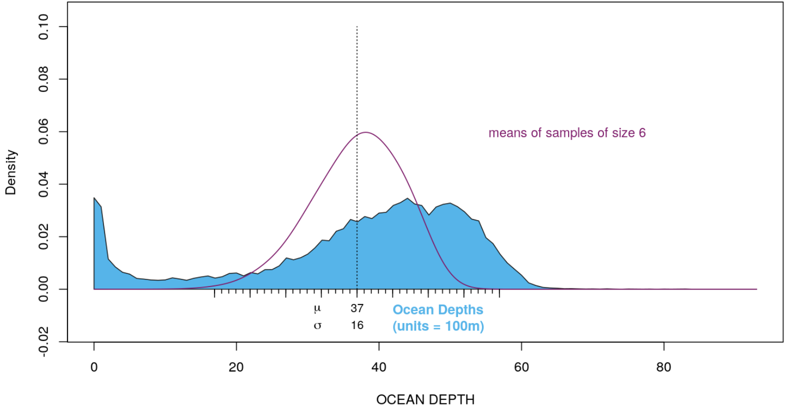
CLT in action: Depths of the ocean



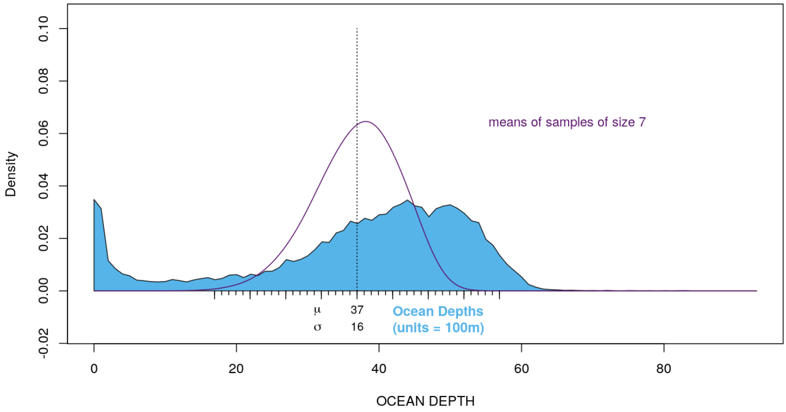
CLT in action: Depths of the ocean



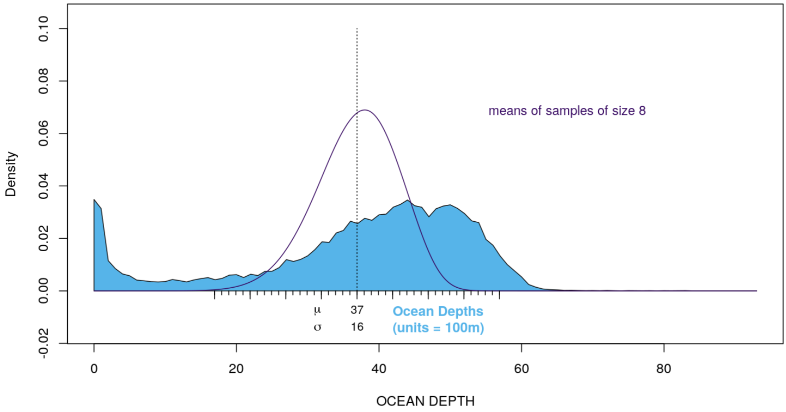
CLT in action: Depths of the ocean



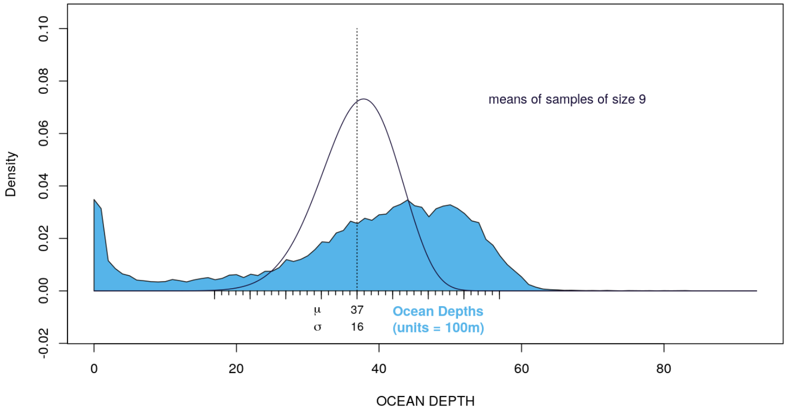
CLT in action: Depths of the ocean



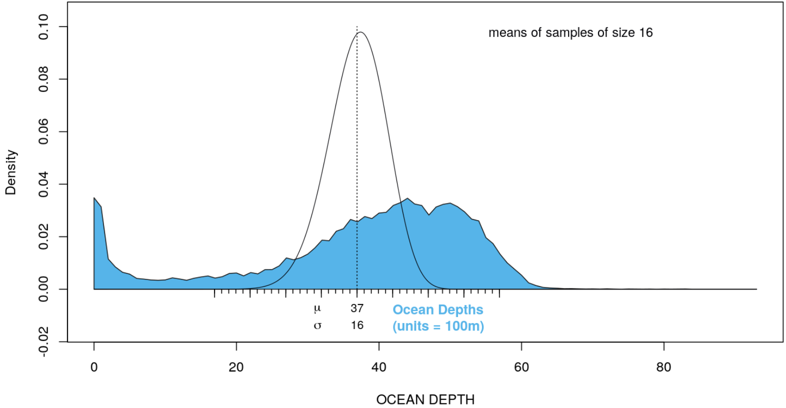
CLT in action: Depths of the ocean



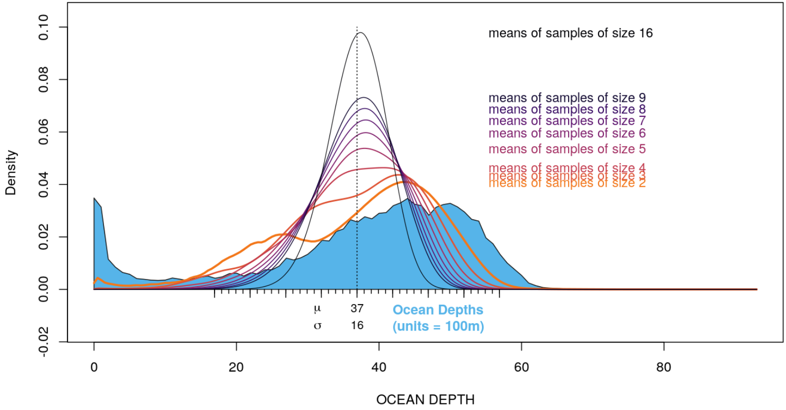
CLT in action: Depths of the ocean



CLT in action: Depths of the ocean

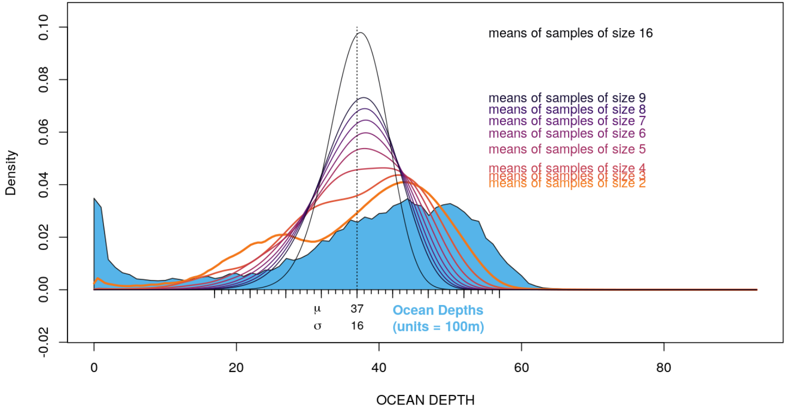


CLT in action: Depths of the ocean



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

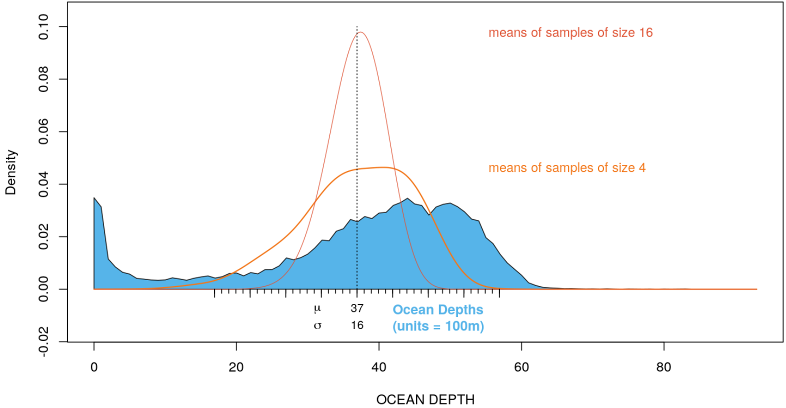


Figure: When the sample size increases from 4 to 16, the spread of the sampling distribution for the mean is reduced by a half, i.e., the range is cut in half. This is known as the curse of the $\sqrt{n}$

Session Info

```
R version 4.0.4 (2021-02-15)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Pop!_OS 21.04

Matrix products: default
BLAS:   /usr/lib/x86_64-linux-gnu/openblas-pthread/libblas.so.3
LAPACK: /usr/lib/x86_64-linux-gnu/openblas-pthread/libopenblasp-r0.3.13.so

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

other attached packages:
 [1] DT_0.16          mosaic_1.7.0     Matrix_1.3-2     mosaicData_0.20.1
 [5] ggformula_0.9.4  ggstance_0.3.4   lattice_0.20-41  kableExtra_1.2.1
 [9] socviz_1.2       gapminder_0.3.0  here_0.1         NCStats_0.4.7
[13] FSA_0.8.30       forcats_0.5.1    stringr_1.4.0    dplyr_1.0.7
[17] purrr_0.3.4      readr_1.4.0      tidyr_1.1.3      tibble_3.1.3
[21] ggplot2_3.3.5    tidyverse_1.3.0  knitr_1.33

loaded via a namespace (and not attached):
 [1] fs_1.5.0           lubridate_1.7.9    webshot_0.5.2      httr_1.4.2
 [5] rprojroot_2.0.2    backports_1.2.1    utf8_1.2.2         R6_2.5.1
 [9] DBI_1.1.1          colorspace_2.0-2   withr_2.4.2        tidyselect_1.1.1
[13] gridExtra_2.3      leaflet_2.0.3      curl_4.3.2         compiler_4.0.4
[17] cli_3.0.1          rvest_1.0.0        pacman_0.5.1       xml2_1.3.2
[21] ggdendro_0.1.22    mosaicCore_0.8.0   scales_1.1.1       digest_0.6.27
[25] foreign_0.8-81     rmarkdown_2.9.7    rio_0.5.16         pkgconfig_2.0.3
[29] htmltools_0.5.2    highr_0.9          dbplyr_1.4.4       fastmap_1.1.0
[33] htmlwidgets_1.5.3  rlang_0.4.11       readxl_1.3.1       rstudioapi_0.13
[37] farver_2.1.0       generics_0.1.0     jsonlite_1.7.2     crosstalk_1.1.1
[41] zip_2.2.0          car_3.0-9          magrittr_2.0.1     Rcpp_1.0.7
[45] munsell_0.5.0      fansi_0.5.0        abind_1.4-5        lifecycle_1.0.0
[49] stringi_1.7.3      carData_3.0-4      MASS_7.3-53.1      plyr_1.8.6
[53] grid_4.0.4         blob_1.2.1         ggrepel_0.8.2      crayon_1.4.1
[57] cowplot_1.1.0      haven_2.3.1        splines_4.0.4      hms_1.0.0
[61] pillar_1.6.2       reprex_0.3.0       glue_1.4.2         evaluate_0.14
[65] data.table_1.14.0  modelr_0.1.8       vctrs_0.3.8        tweenr_1.0.1
[69] cellranger_1.1.0   gtable_0.3.0       polyclip_1.10-0    assertthat_0.2.1
[73] TeachingDemos_2.12 xfun_0.25          ggforce_0.3.2      openxlsx_4.1.5
[77] broom_0.7.2        viridisLite_0.4.0  ellipsis_0.3.2
```