Topic 2: Sampling distribution of the mean

Background for inference of location parameter in location/scale families

Sonja Petrović
Created for ITMD/ITMS/STAT 514

Spring 2021.

Context

Statistics:

- Functions of random variables
- Therefore, are random variables themselves.
  - In particular, they have their own distributions, called sampling distributions.
  - Meaning of sampling distribution **(Review)**
- How does inference relate to analytics? **(Review)**

Context

Statistics:

- Functions of random variables
- Therefore, are random variables themselves.
  - In particular, they have their own distributions, called sampling distributions.
  - Meaning of sampling distribution **(Review)**
- How does inference relate to analytics? **(Review)**

Sampling distribution of the mean

CLT

If $\overline{X}$ is the mean of a random sample of size n taken from a population with mean μ and finite variance σ^2, then the limiting form (as $n \to \infty$) of the distribution of the random variable

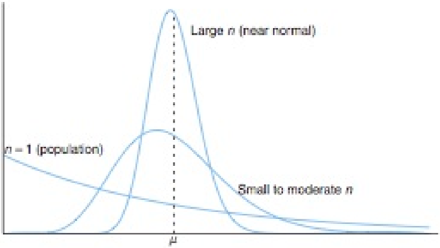
$$Z := \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

is the stanard normal distribution $N(0, 1)$.

$$\mu_{\overline{X}} = \mu_{population}; \quad \sigma_{\overline{X}} = \sigma_{population}/\sqrt{n}.$$

Sampling distribution of the mean, part 2

- The presumption of normality on the distribution of $\overline{X}$ becomes more accurate as n grows larger.



- How the distribution of $\overline{X}$ becomes closer to normal as n grows larger, beginning with the clearly nonsymmetric distribution of an individual observation ($n = 1$).
- The mean of $\overline{X}$ remains μ for any sample size and the variance of $\overline{X}$ gets smaller as n increases.

Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution

The sampling distribuion of $\overline{X}$ is ...

Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution

The sampling distribuion of $\overline{X}$ is ... approximately normal with $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = 40/\sqrt{16} = 10$.

Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution

The sampling distribuion of $\overline{X}$ is ... approximately normal with $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = 40/\sqrt{16} = 10$.
Corresponding to $\overline{X} = 775$, the observed Z-value is

...

Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution

The sampling distribuion of $\overline{X}$ is ... approximately normal with $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = 40/\sqrt{16} = 10$.
Corresponding to $\overline{X} = 775$, the observed Z-value is ...$(775 - 800)/10 = -2.5$, and therefore

Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

Solution

The sampling distribuion of $\overline{X}$ is ... approximately normal with $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = 40/\sqrt{16} = 10$.
Corresponding to $\overline{X} = 775$, the observed Z-value is ...$(775 - 800)/10 = -2.5$, and therefore

$$P(\overline{X} < 775) = P(Z < -2.5) = 0.02275$$

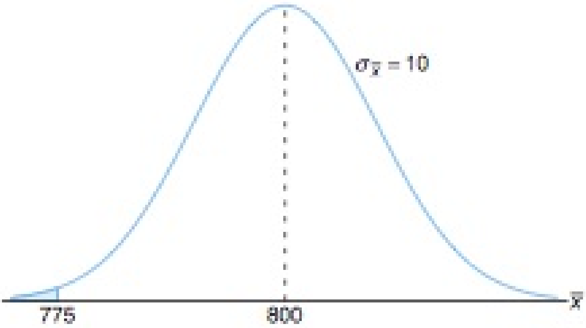
Example of use of CLT

An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. *Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.*

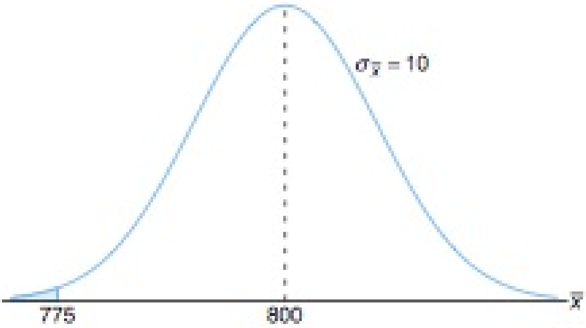
Solution

The sampling distribuion of $\overline{X}$ is ... approximately normal with $\mu_{\overline{X}} = 800$ and $\sigma_{\overline{X}} = 40/\sqrt{16} = 10$.
Corresponding to $\overline{X} = 775$, the observed Z-value is ...$(775 - 800)/10 = -2.5$, and therefore

$$P(\overline{X} < 775) = P(Z < -2.5) = 0.02275$$



The desired probability is given by the area of the shaded region in the figure.



The desired probability is given by the area of the shaded region in the figure.

Aha!

Discuss the meaning: ahaslides.com/STATITMW4

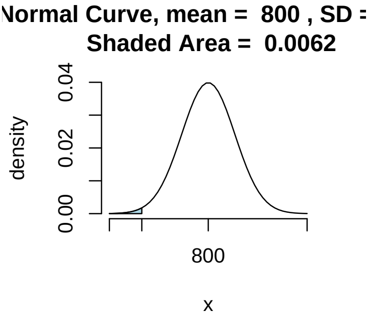
Example: solving the exact same problem in R

We can use basic R command “pnorm”, or a graphical version of it from the package “tigerstats” to help us plot:

```
pnorm(775,mean=800,sd=10)
```

```
[1] 0.006209665
```

```
pnormGC(775, region="below", mean=800, sd=10,graph=TRUE)
```



Example: solving the exact same problem in Python

The Python command that is equivalent to R's "pnorm" is called "norm.cdf".

```
from scipy.stats import norm
norm.cdf(775,loc=800,scale=10)
```

```
0.006209665325776132
```

The "p" in the name "pnorm" refers to *probability*, and the "cdf" in Python to *cumulative density function*. Both refer to the same statistical concept: the command called for the number 775 will compute the probability of a normally distributed random variable to be ≤ 775. Which normal distribution? well, that's why we give mean (or loc) and stdev (or scale) as parameters to these functions.

Case study 1: automobile parts

Problem.

An important manufacturing process produces cylindrical component parts for the automotive industry. It is important that the process produce parts having a mean diameter of 5.0 millimeters. The engineer involved claims that the population mean is 5.0 millimeters.

An experiment is conducted in which 100 parts produced by the process are selected randomly and the diameter measured on each. It is known that the population standard deviation is $\sigma = 0.1$ millimeter. The experiment indicates a sample average diameter of $\overline{X} = 5.027$ millimeters.

Question:

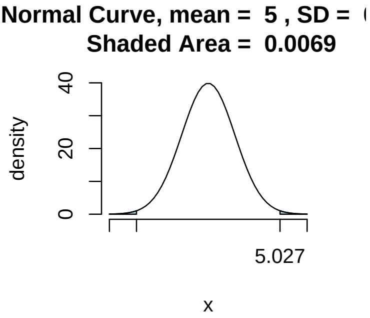
Does this sample information appear to support or refute the engineer's claim?

Case study 1: Solution

Having spent some time with the case study, you should arrive at this solution:

The probability that we choose to compute is given by:

$$P(|\overline{X}-5| \geq 0.027) = 2P\left(\frac{\overline{X}-5}{0.1\sqrt{100}} \geq 2.7\right) = 0.0035 = 0.007$$




```
[1] 0.006933948
```

What is next?

A similar result holds for a difference in two means. We will see:

- an example,
- the theorem, and
- a hands-on case study.

License