Distribution of Sample Means

- We say that the sample mean ($\bar{y}$) gives us *an estimate* of the population mean µ
- But, since the sample is only a small subset of the entire population, *$\bar{y}$ is an uncertain estimate of μ*

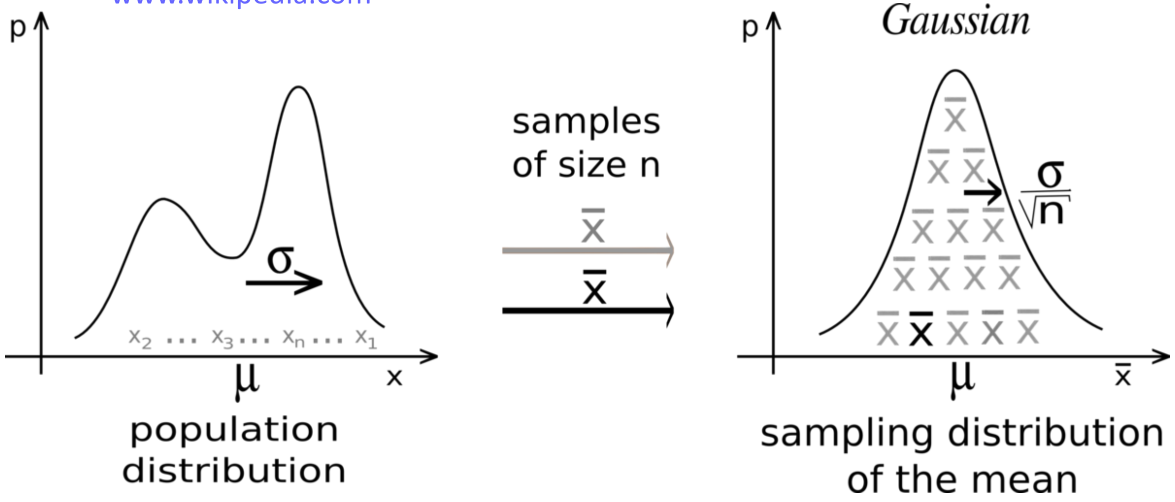
- What if, we sample the population *several times*, each time calculating the sample mean $\bar{y}$
- Let's say we do it 'k' times, we get sample means as $\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \ldots.., \bar{y}_k$
- How would these sample means behave?
  - How close are they from *μ?*
  - What's the *mean of sample means*?
  - What's the *standard deviation of sample means*?
  - More importantly, *what's their frequency distribution*?

Central Limit Theorem

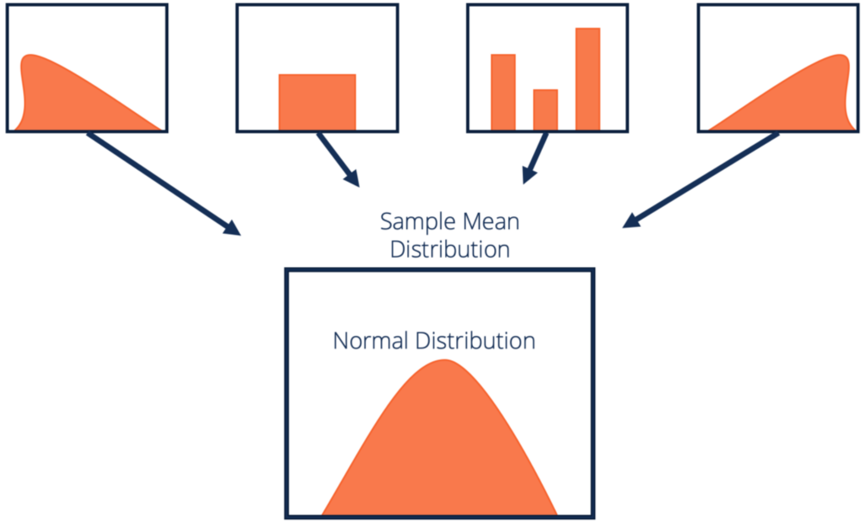
"The *distribution of sample means* ($\bar{y}_1, \bar{y}_2, \bar{y}_3, \bar{y}_4, \ldots.., \bar{y}_k$) follows *a normal distribution*, even when the original variable y is NOT normally distributed."

www.moriah.com

www.wikipedia.com



p
x_2 ... x_3 ... x_n ... x_1
σ
μ
x
population distribution
samples of size n
$\bar{x}$
$\bar{x}$
Gaussian
p
$\frac{\sigma}{\sqrt{n}}$
μ
$\bar{x}$
sampling distribution of the mean



- What is the mean of distribution of sample means?

- What is the variance of distribution of sample means?

NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

Central Limit Theorem

DIY: Read

“In non-mathematical language, the “CLT” says that whatever the PDF of a variable is, if we randomly sample a “large” number (say k) of independent values from that random variable, the sum or mean of those k values, if collected repeatedly, will have a Normal distribution.

It takes some extra thought to understand what is going on here. The process I am describing here takes a sample of (independent) outcomes, e.g., the weights of all of the rats chosen for an experiment, and calculates the mean weight (or sum of weights). Then we consider the less practical process of repeating the whole experiment many, many times (taking a new sample of rats each time). If we would do this, the CLT says that a histogram of all of these mean weights across all of these experiments would show a Gaussian shape, even if the histogram of the individual weights of any one experiment were not following a Gaussian distribution.

By the way, the distribution of the means across many experiments is usually called the sampling distribution of the mean.”

- Seltman, Howard J. "Experimental design and analysis." (2012)

Distribution of Sample Means

- The Central Limit Theorem is the explanation for why many real-world random variables tend to have a Gaussian distribution. It is also the justification for assuming that if we could repeat an experiment many times, any sample mean that we calculate once per experiment would follow a Gaussian distribution over the many experiments.
- Since the *distribution of the sample means* with mean (μ) and variance (σ_y^2/n) follows a normal distribution, then the relationship between the distribution of sample means and the z-distribution is given by:

$$z = \frac{\bar{y} - \mu}{\frac{\sigma_y}{\sqrt{n}}}$$

- What does it tell about value of a random sample mean?
- But, we often don’t know the population standard deviation (σ_y) or variance (!)
- **Can we estimate them?**

Estimators of Population

Ref: Montgomery, D. C., "Design and Analysis of Experiments," 8th Ed.,

We may easily show that $\bar{y}$ and S^2 are unbiased estimators of μ and σ^2, respectively.

First consider $\bar{y}$. Using the properties of expectation, we have

$$
\begin{aligned}
E(\bar{y}) &= E\left(\frac{\sum_{i=1}^{n} y_i}{n}\right) \\
&= \frac{1}{n} \sum_{i=1}^{n} E(y_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \mu \\
&= \mu
\end{aligned}
$$

$$
\begin{aligned}
E(S^2) &= E\left[\frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1}\right] \\
&= \frac{1}{n-1} E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] \\
&= \frac{1}{n-1} E(SS)
\end{aligned}
$$

$$
\begin{aligned}
E(SS) &= E\left[\sum_{i=1}^{n} (y_i - \bar{y})^2\right] \\
&= E\left[\sum_{i=1}^{n} y_i^2 - n\bar{y}^2\right] \\
&= \sum_{i=1}^{n} (\mu^2 + \sigma^2) - n(\mu^2 + \sigma^2/n) \\
&= (n-1)\sigma^2
\end{aligned}
$$

$$
E(S^2) = \frac{1}{n-1} E(SS) = \sigma^2
$$

S^2 is an unbiased estimator of σ^2.

where $SS = \Sigma_{i=1}^{n}(y_i - \bar{y})^2$ is the **corrected sum of squares** of the observations y_i.

Degrees of Freedom (DOF)

If y is a random variable with variance σ^2 ,

and sum of squares SS = $\sum(y_i - \bar{y})^2$ has ‘v’ degrees of freedom, then $E\left(\frac{SS}{v}\right) = \sigma^2$

The number of **degrees of freedom of a sum of squares** is equal to the number of independent elements in that sum of squares.

For example, SS = $\sum(y_i - \bar{y})^2$ is a sum of squares of ‘n’ elements, i.e., $y_1 - \bar{y}$, $y_2 - \bar{y}$, …, $y_n - \bar{y}$

Note that these ‘n’ elements are not all independent because $\sum(y_i - \bar{y}) = 0$

Therefore, only n-1 of them are independent, implying that SS has (n-1) degrees of freedom.

$$E\left(\frac{SS}{n-1}\right) = \sigma^2$$

NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

Consequence of CLT

If $y_1, y_2, y_3 \ldots, y_n$ is a sequence of ‘n’ independent and identically distributed random variables

with $E(y_i) = \mu$ and $\mathrm{V}(y_i) = \sigma^2$ (both finite)

If we define, $x = y_1 + y_2 + y_3 + \cdots + y_n$ Then what is the distribution of ‘x’ as ‘n’ becomes sufficiently large?

$$= n\,\bar{y}$$

In other words, $Z_n = \frac{x - n\mu}{\sqrt{n\sigma^2}}$ is a standard normal distribution as $n \to \infty$

IMP!

‘sum of n independent and identically distributed random variables is approximately normally distributed’

Frequently, we think of the error in an experiment as arising in an additive manner from several independent sources; consequently, the normal distribution becomes a reasonable model for the combined experimental error.

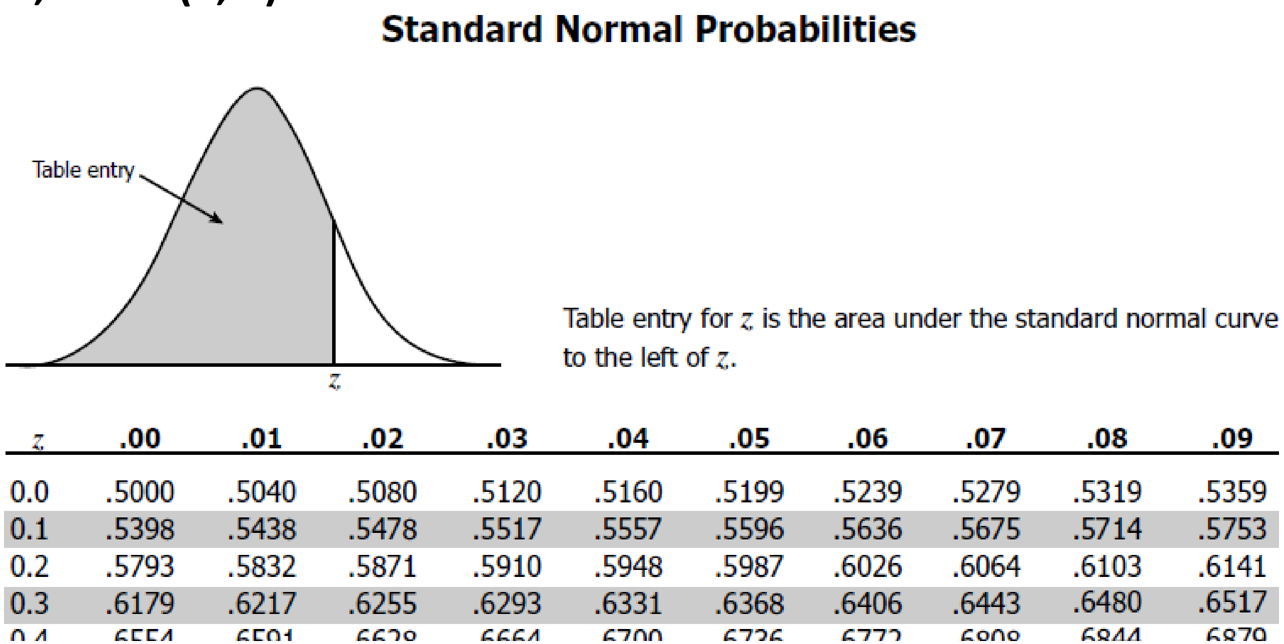
NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.

z-Distribution (Std. Normal PDF)

If $y_1, y_2, \ . \ . \ . \ , y_n$ is a random sample from the **ANY** distribution, then

$$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$

is distributed as **Standard Normal Distribution, i.e., NPDF (0, 1)**



Standard Normal Probabilities

Table entry
z

Table entry for z is the area under the standard normal curve to the left of z.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879

DIY Read Chapter 2 Pages 32- 52 from Design and Analysis of Experiments, 8th Ed.

NOTE: You do NOT have permission to share this file or any of its contents with anyone else, and/or upload it on internet or any of the platforms where it can be accessed by others.