Chapter 10 - The Gaussian Distribution

Laura Saba

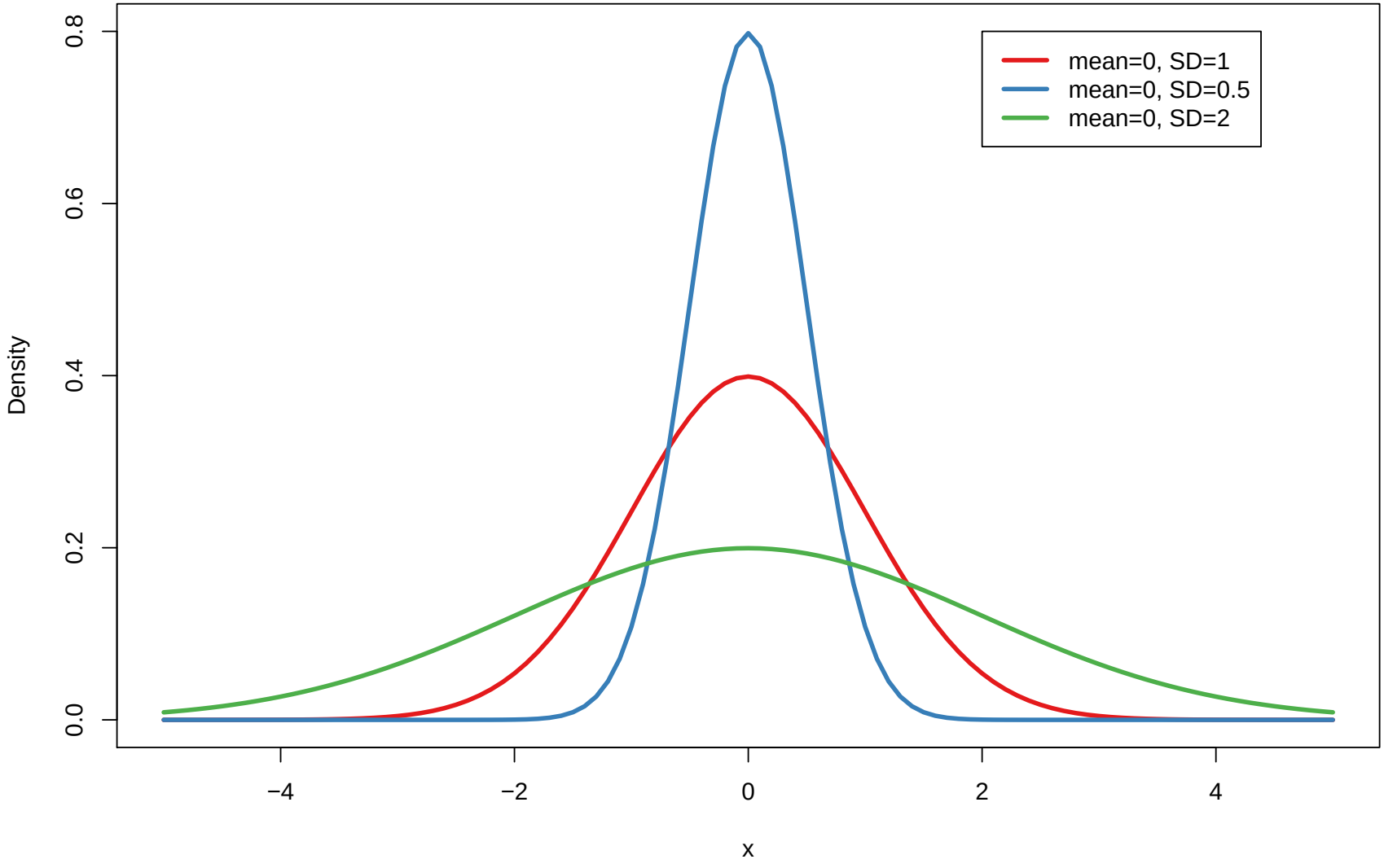
July 11, 2017

What this chapter covers

- The nature of the Gaussian distribution
- SD and the Gaussian distribution
- The standard normal distribution
- The normal distribution does not define normal limits
- Why the Gaussian distribution is so central to statistical theory

THE NATURE OF THE GAUSSIAN DISTRIBUTION

Symmetrical Bell-Shapped Distribution



Many random factors

When many (independent) random factor contribute to an observed value, the observed values tend to follow Gaussian distribution.

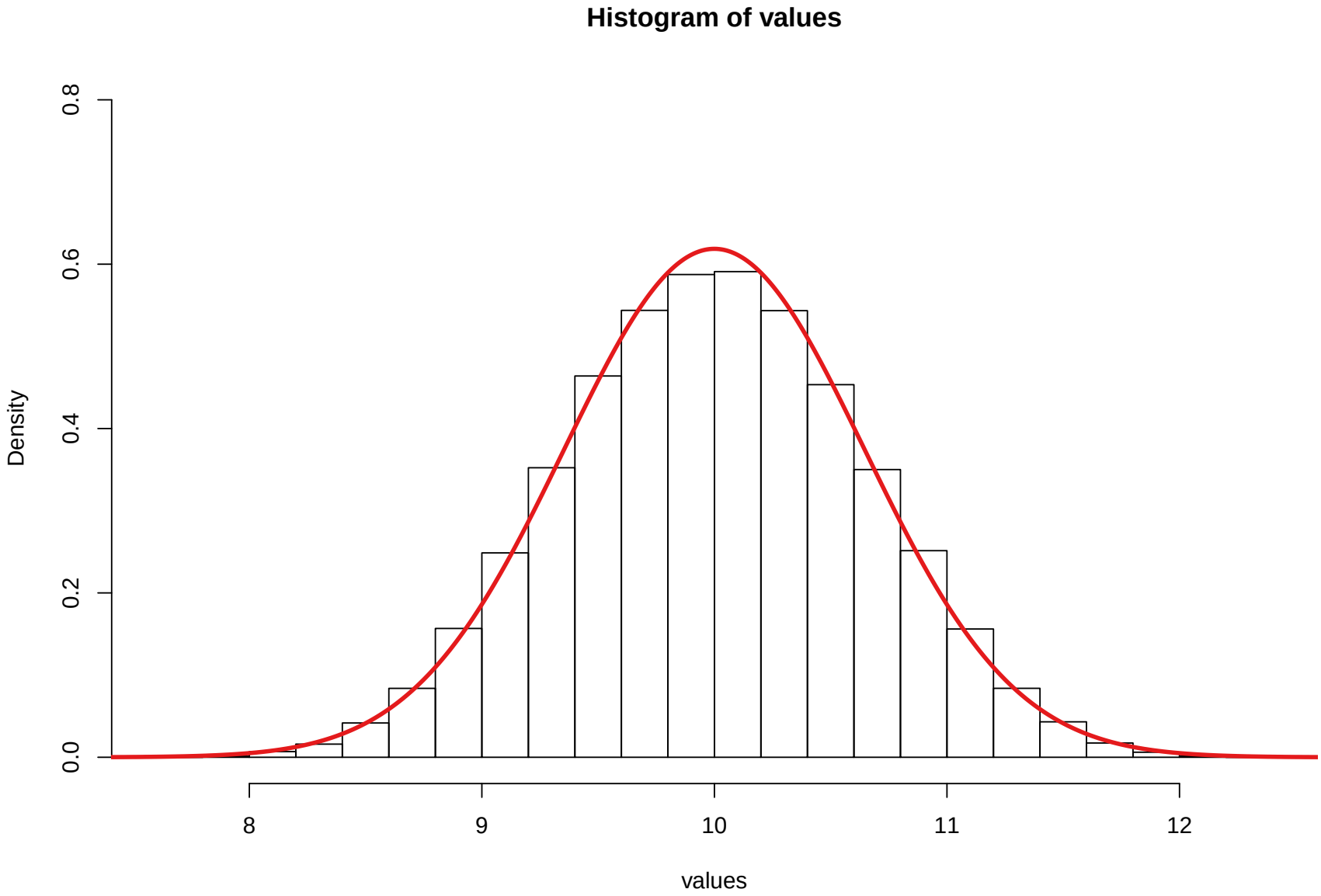
```
true_value = 10
e1 = runif(100000) - 0.5
e2 = runif(100000) - 0.5
e3 = runif(100000) - 0.5
e4 = runif(100000) - 0.5
e5 = runif(100000) - 0.5

values = true_value + e1 + e2 + e3 + e4 + e5
head(round(values,2))
```

```
## [1] 10.29  9.57  9.61  9.32 11.02  9.95
```

```
tail(round(values,2))
```

```
## [1]  9.10 11.50 10.28  9.36 10.64  9.41
```

Result of many random factors contributing



Common Statistical Tests and the Gaussian Distribution

Many commonly used statistical test rely on the assumption that the data being sampled from a population that follows a Gaussian distribution.

This is often a reasonable assumption.

SD AND THE GAUSSIAN DISTRIBUTION

Ideal Gaussian Distribution In Terms of SD

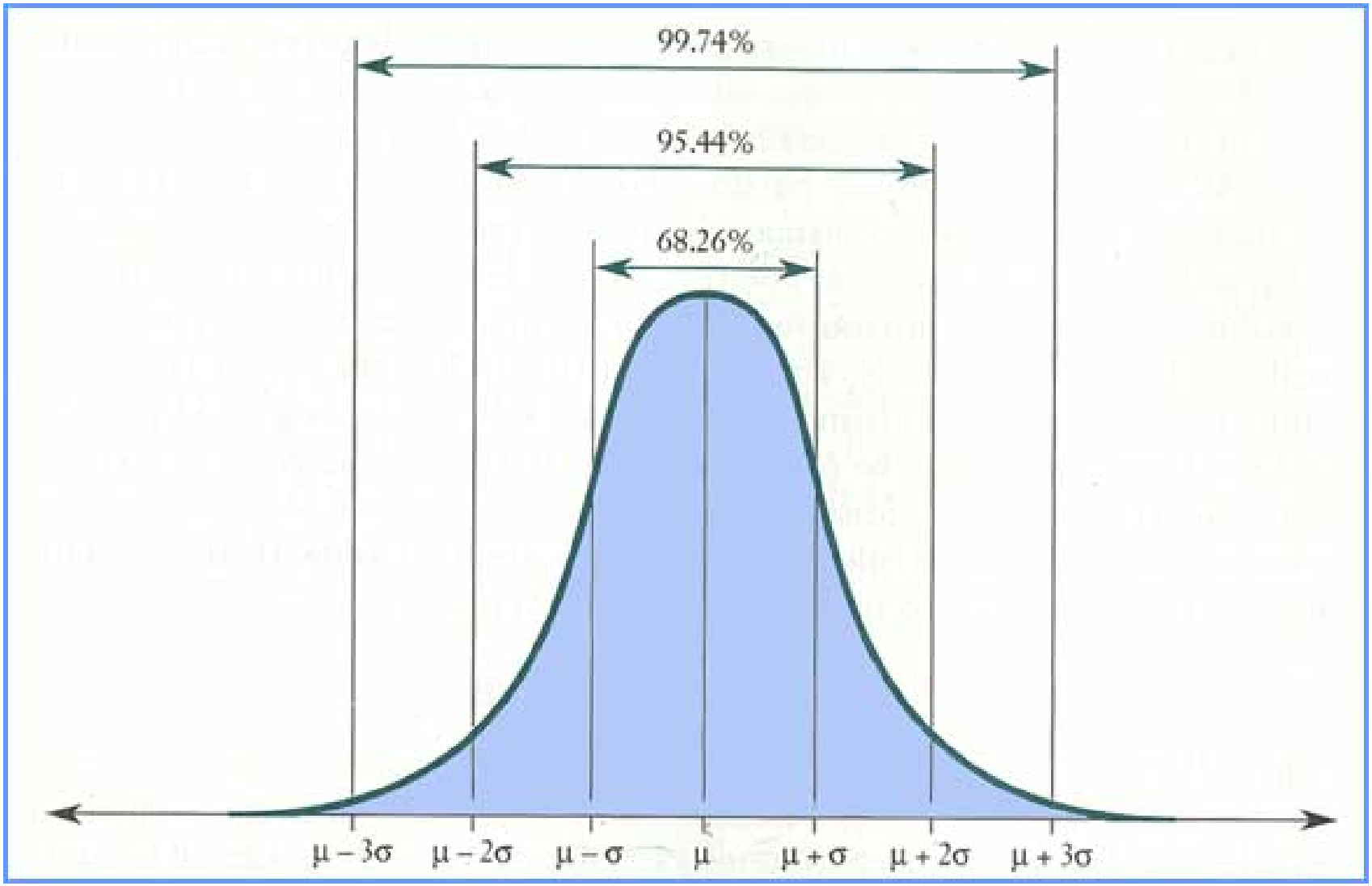


Figure 1: Ideal Gaussian Distribution

How well does the data fit this rule

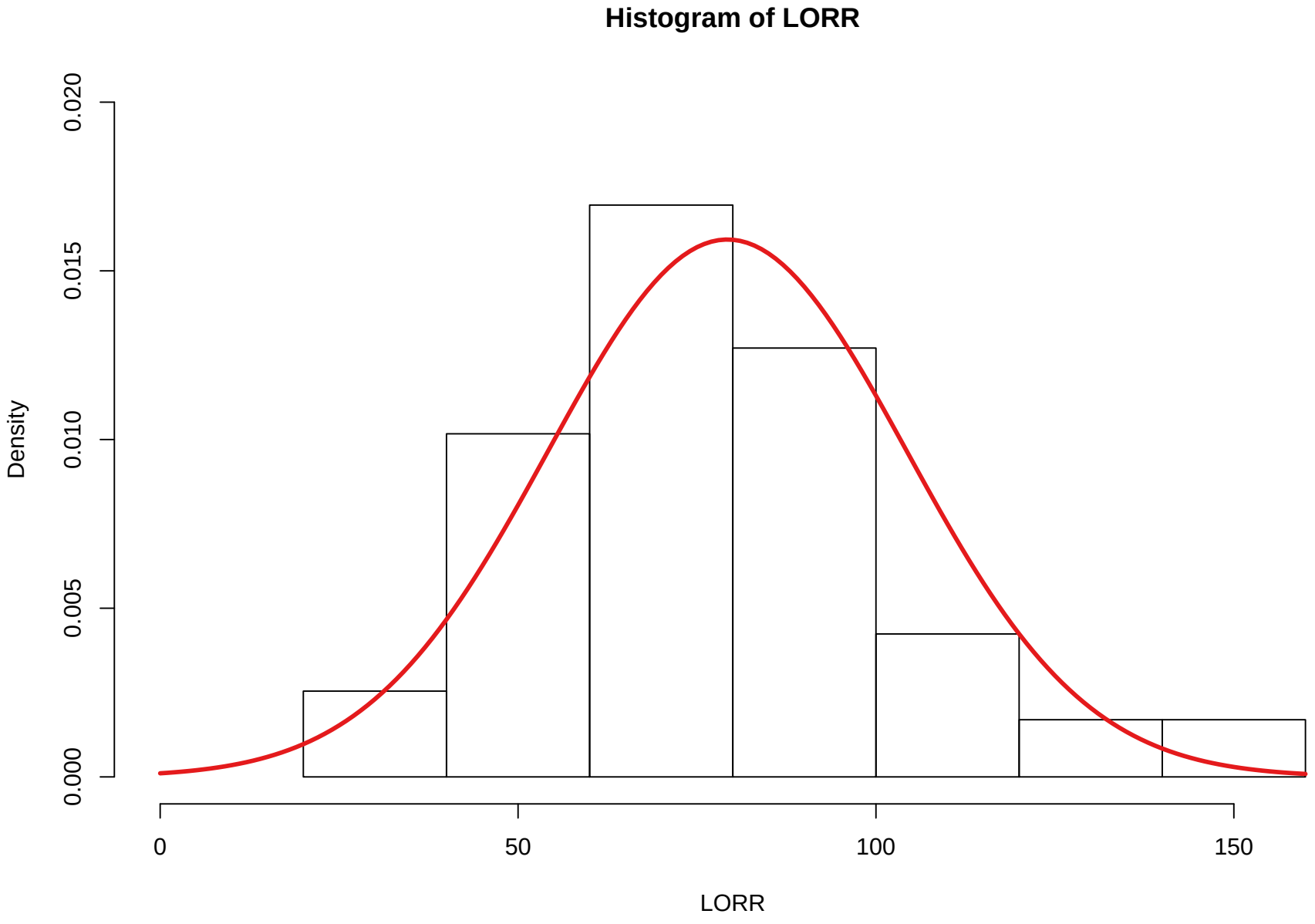
```
n = length(values)
sd1=sum(abs(values-mean(values))<(1*sd(values)))/n
sd2=sum(abs(values-mean(values))<(2*sd(values)))/n
sd3=sum(abs(values-mean(values))<(3*sd(values)))/n
```

1 SD: simulated data set = 67.21% theoretical = 68.26%

2 SD: simulated data set = 95.75% theoretical = 95.44%

3 SD: simulated data set = 99.90% theoretical = 99.74%

How well does the Gaussian distribution fit the LORR data



How well does the Gaussian distribution fit the LORR data

```
n = length(LORR)
st1=sum(abs(LORR-mean(LORR))<(1*sd(LORR)))/n
st2=sum(abs(LORR-mean(LORR))<(2*sd(LORR)))/n
st3=sum(abs(LORR-mean(LORR))<(3*sd(LORR)))/n
```

1 SD: simulated data set = 72.88% theoretical = 68.26%

2 SD: simulated data set = 94.92% theoretical = 95.44%

3 SD: simulated data set = 100.00% theoretical = 99.74%

THE STANDARD NORMAL

Standard Normal - Gaussian distribution with mean = 0 and SD = 1

All Gaussian distributions can be converted to a standard normal distribution using the following formula:

$z = \frac{Value - Mean}{SD}$

z is the number of SD the value is away from the mean

The standard normal distribution

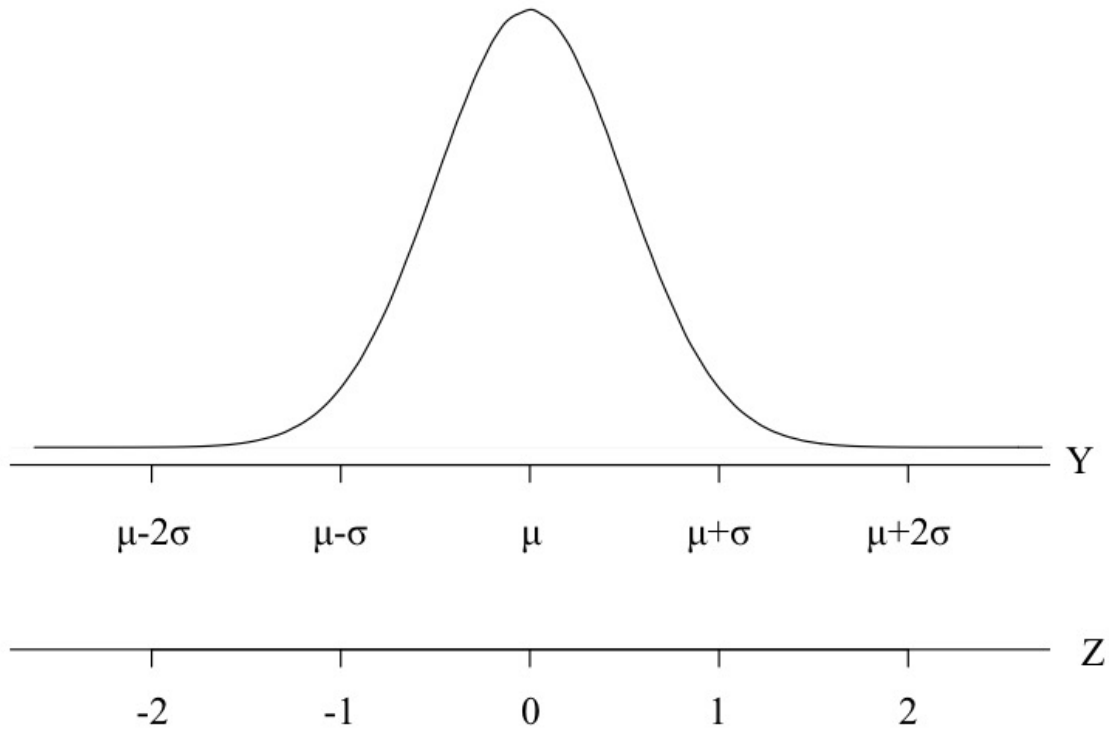


Figure 2: The Standard Normal Distribution

The standard normal distribution

Table 10.1 The standard normal distribution.

z	PERCENTAGE OF STANDARD NORMAL DISTRIBUTION BETWEEN -z AND z
0.67	50.00
0.97	66.66
1.00	68.27
1.64	90.00
1.96	95.00
2.00	95.45
2.58	99.00
3.00	99.73

Calculate Table 10.1 in R

```
z = c(0.67,0.97,1.0,1.65,1.96,2,2.58,3)
pct = 1-pnorm(z,lower.tail = FALSE)*2
round(pct*100,2)
```

```
## [1] 49.71 66.80 68.27 90.11 95.00 95.45 99.01 99.73
```

```
## Taking z to 4 significant digits
z = c(0.6745,0.9673,1,1.645,1.96,2,2.576,3)
pct = 1-pnorm(z,lower.tail = FALSE)*2
round(pct*100,2)
```

```
## [1] 50.00 66.66 68.27 90.00 95.00 95.45 99.00 99.73
```

THE NORMAL DISTRIBUTION DOES NOT DEFINE NORMAL LIMITS

- Normal distribution does not equate to 'normal' range
- Defining the normal limits of a clinical measurement is not straightforward and requires clinical thinking, not just statistics

WHY THE GAUSSIAN DISTRIBUTION IS SO CENTRAL TO STATISTICAL THEORY

- The Central Limit Theorem states that regardless of the distribution of the population, if the number of observations is large enough (typically 30 or greater) , then the sampling distribution of the sample mean is at least APPROXIMATELY normal (i.e., Gaussian).
- This is an extremely important theorem! It allows us to use statistical inference based on the normal distribution for statistics where the sample size is relatively large.

Sampling Distribution of the Sample Mean

- **Sampling variation** is the concept that (random) samples from the same population will differ because they contain different members of the population.
- **Sampling distribution** is the frequency distribution of the statistics resulting from all possible samples (of a certain size n) from the same population.
- Example: Suppose we could get ratings of all statistics instructors in the world and that the population mean rating (on a scale from 1 to 20) is μ=10. We take several random samples of size n=4 and calculate the sample mean. What is the distribution of the sample means?

Sampling Distribution of the Sample Mean

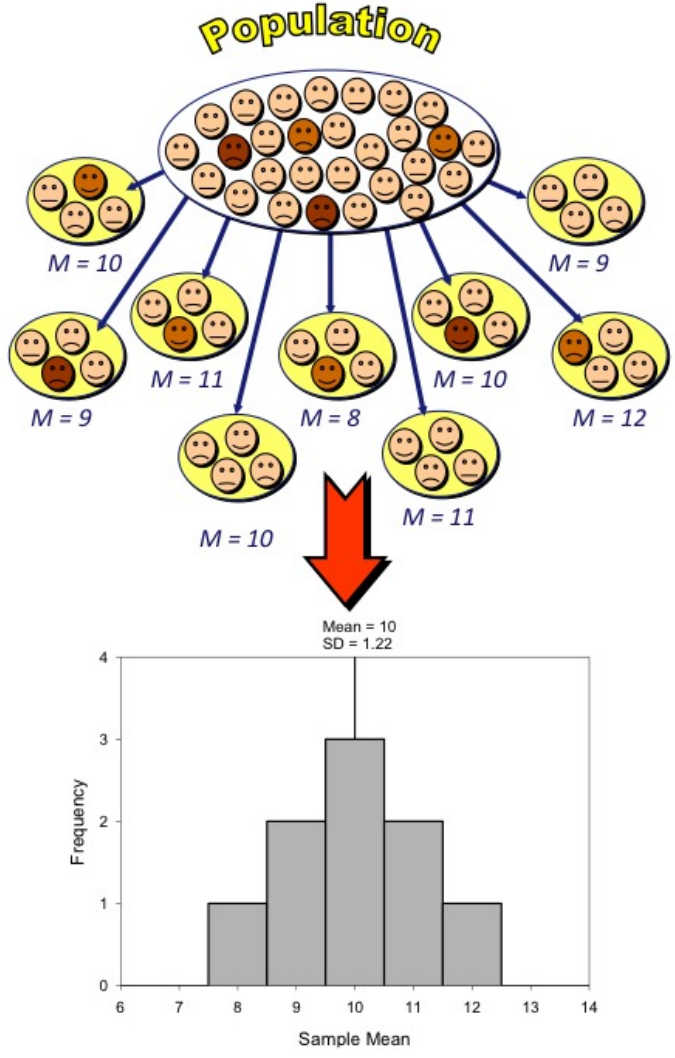


Figure 3: Sampling Distribution

What did we learn

- The Gaussian bell-shaped distribution is the basis for much of statistics. It arises when many random factors create variability.
- The Gaussian distribution is also called a normal distribution. But this use of normal is very different than the usual use of that word to mean ordinary or abundant.
- The central limit theorem explains why Gaussian distributions are central to much of statistics.