Central Limit Theorem

Week 02 - Day 02

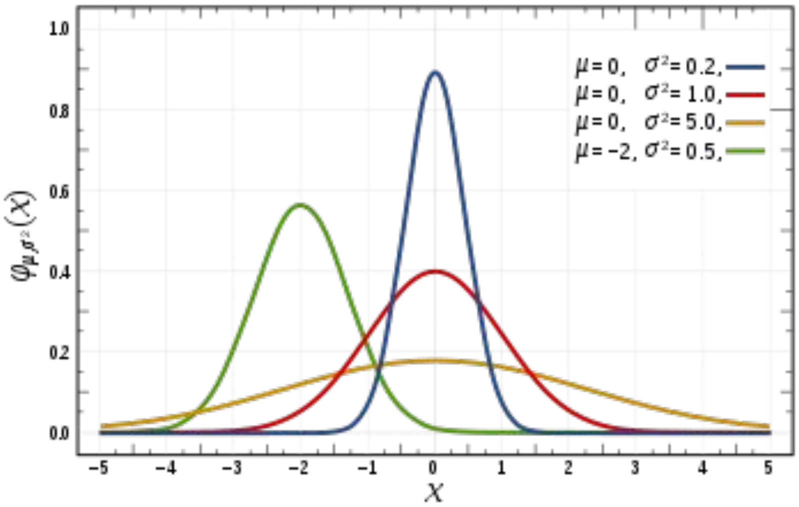
Normal Distribution

Names

Normal distribution

Gaussian distribution

Bell-shaped distribution




GAUSSIAN
GAUSSIAN IS EVERYWHERE
makeameme.or

Examples

1. Age of all the students in GA

Examples

1. Age of all the students in GA
2. Height of all the trees in a forest

Examples

1. Age of all the students in GA
2. Height of all the trees in a forest
3. Number of grains of rice eaten by ppl in Singapore every year

Examples

1. Age of all the students in GA
2. Height of all the trees in a forest
3. Number of grains of rice eaten by ppl in Singapore every year
    a. **bimodal if you add Europe!**

Parameters

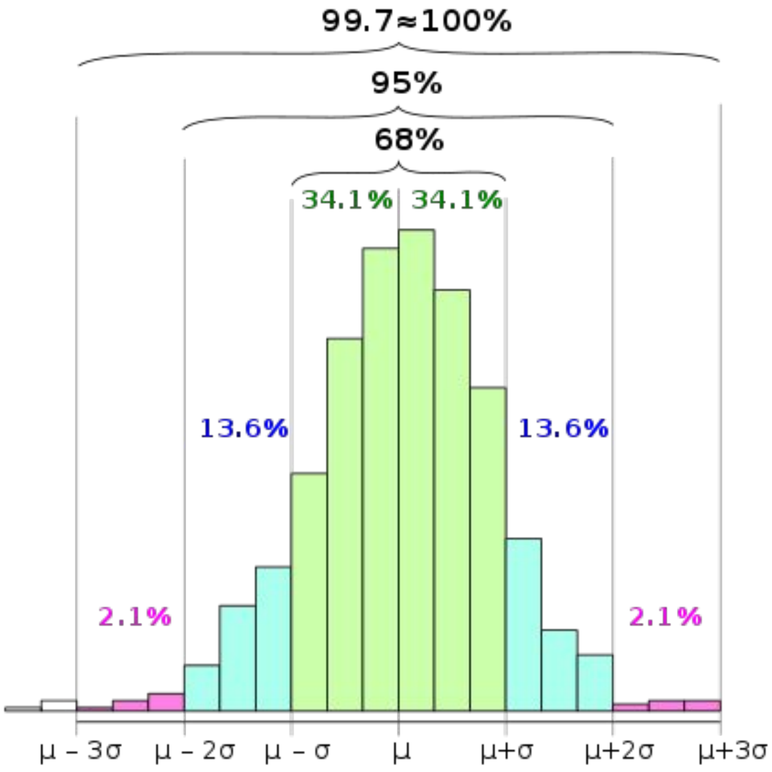
1. Mean
2. STD

Mean

numpy.random.normal(*loc=100*, *scale=15*, *size=1000*)

STD

The 68-95-99.7 rule



Standard Normal distribution

Mean = 0

Std = 1

x = 3.5 -> “3.5 std(s) far from the mean”

The z-score

$Z = (x - X\_mean) / X\_std$

- Normal distribution ➜ Standard normal distribution
- Used is ML!

Central Limit Theorem

Population vs. Sample

Metric	Statistic	Parameter
mean	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum x}{N}$
standard deviation	$s = \sqrt{\frac{\sum_i (x_i - \bar{x})^2}{n-1}}$	$\sigma = \sqrt{\frac{\sum_i (x_i - \mu)^2}{N}}$
correlation	$r = \frac{\hat{Cov}(X, Y)}{s_X s_Y}$	$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y}$

“The sampling distribution of X_mean is normally distributed…even if the distribution of is not!”

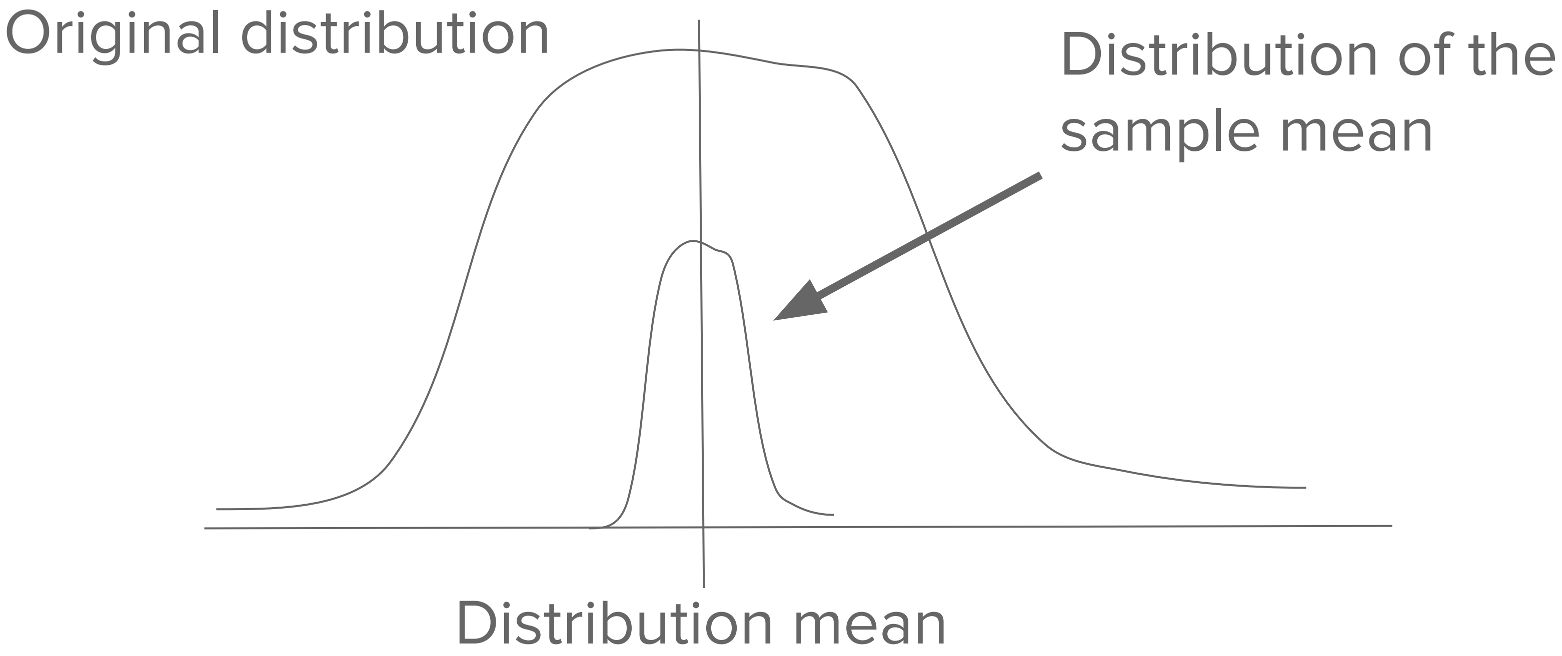

SO WHATS
YOUR POINT?

Property 1 from CLT

If $X \sim N(\mu, \sigma)$, then $\bar{X}$ is exactly $N(\mu, \frac{\sigma}{\sqrt{n}})$

Len of the sample

[mean(sample_1), mean(sample_2),...mean(sample_i)]


Original distribution
Distribution of the
sample mean
Distribution mean

Property 2 from CLT

If $\bar{x}$ is normally distributed,

then we can use inferential methods

that rely on our sample mean, $\bar{x}$

PDF vs. Histogram

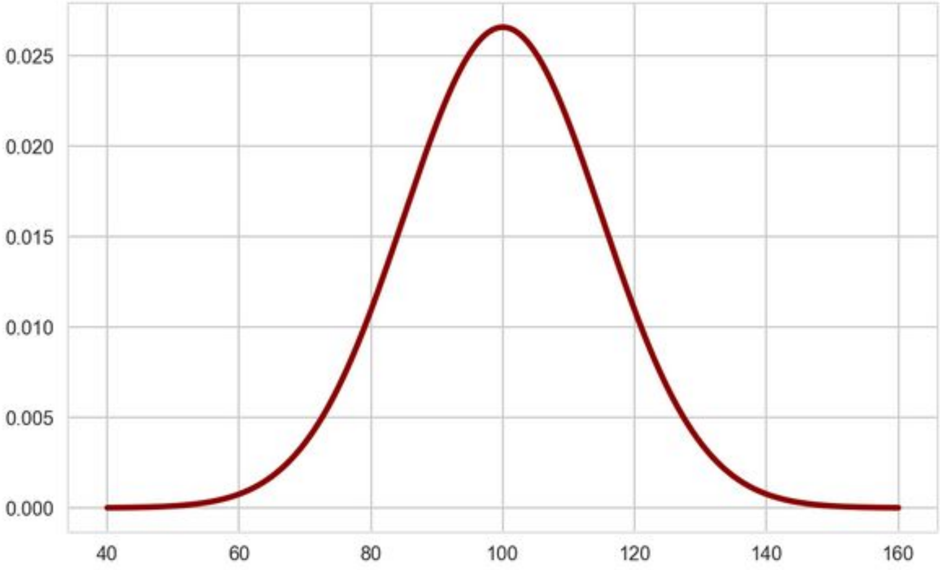
```python
# Generate points on the x axis:
xpoints = np.linspace(40, 160, 500)

# Use stats.norm.pdf to get values on the probability density function
ypoints = stats.norm.pdf(xpoints, 100, 15)

# initialize a matplotlib "figure":
fig, ax = plt.subplots(figsize=(8,5))

# Plot the lines using matplotlib's plot function:
ax.plot(xpoints, ypoints, linewidth=3, color='darkred')
```

```
[<matplotlib.lines.Line2D at 0x10dfb2e50>]
```



PDF vs histogram?

Histogram = count the values in each bin

PDF

sum=1

area between two points

Exercises

1. Play (i.e. plot the histogram, print the percentiles, etc.) with the CLT using a normal distribution
2. Play with CLT using a bimodal distribution (i.e. put together two normal distributions with different means)
3. Play with CLT using a small sample (e.g. n=5)