Statistics MM1: Introduction to statistics

Lecturer: Israel Leyva-Mayorga

email: ilm@es.aau.dk

AALBORG UNIVERSITY
DENMARK

Connectivity

Schedule

1. Introduction to statistics
2. Parameter estimation
3. Confidence intervals
4. Hypothesis testing 1
5. Hypothesis testing 2
6. Regression
7. Workshop: wrap-up and exam problems

Outline

What is statistics?

Classification of statistics

The law of large numbers

The normal (Gaussian) distribution

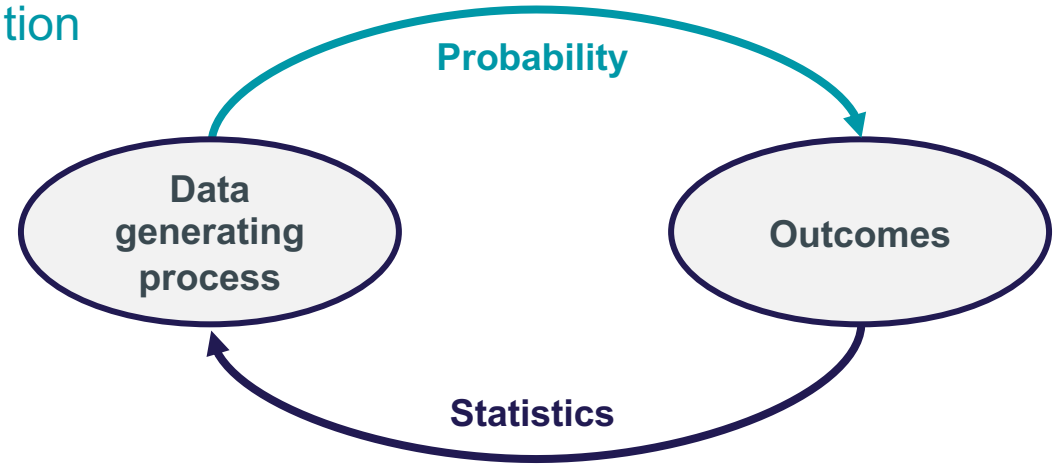
The central limit theorem

What is statistics?

What is statistics?

Probability

Methods that allow us to say something about the possible outcomes of a process from the distribution



Statistics

Methods that allow us to say something about a process from the outcomes

What is statistics?

Methods that allow us to say something about a process from its samples
Basis for data-driven methods and machine learning

Statistical inference:

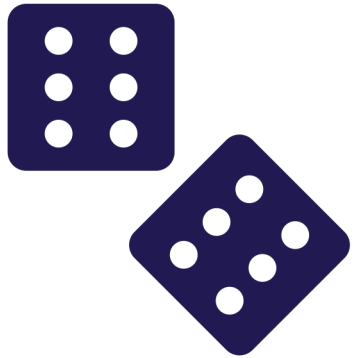
Process of using data samples to infer the distribution that generated them

- Prediction
- Classification
- Clustering
- Estimation

Problem: In everyday statistics, the uncertainty is oftentimes hidden
It's difficult to deal with uncertainty because we are used to live based on certainty

Fair dice example

A fair dice should have the exact same probability of rolling any of the numbers
Can you say if a dice is fair by rolling it once?
Twice?
After rolling it 100 times, how confident are you that it is fair?



Why do we need statistics?

In real life, we rarely know the distribution that generates the data points
We might only have some initial assumptions

Statistics allows us to interpret data

Transform measurements into a model that can be used to make predictions

The scientific method often relies on statistics for empirical confirmation
Hypothesis testing: Can I accept or reject the hypothesis with the data I have?

Classification of statistics

Descriptive and inferential statistics

Descriptive statistics

Used to present the data in an understandable way

How can we show our results to say something meaningful?

From measurements to graphs, plots, charts, …

Inferential statistics

Used to extend the results to a larger population

What can we say about the phenomena from the data we have?

From measurements to predictions

Frequentist and Bayesian statistics

Frequentist statistics

Look only at the data, with fixed parameters

Easy to compute

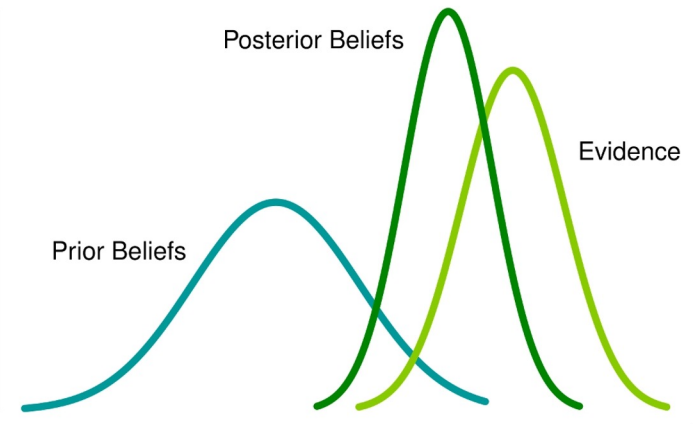
Bayesian statistics

Can include prior knowledge

Can be updated with new information

Parameters are stochastic

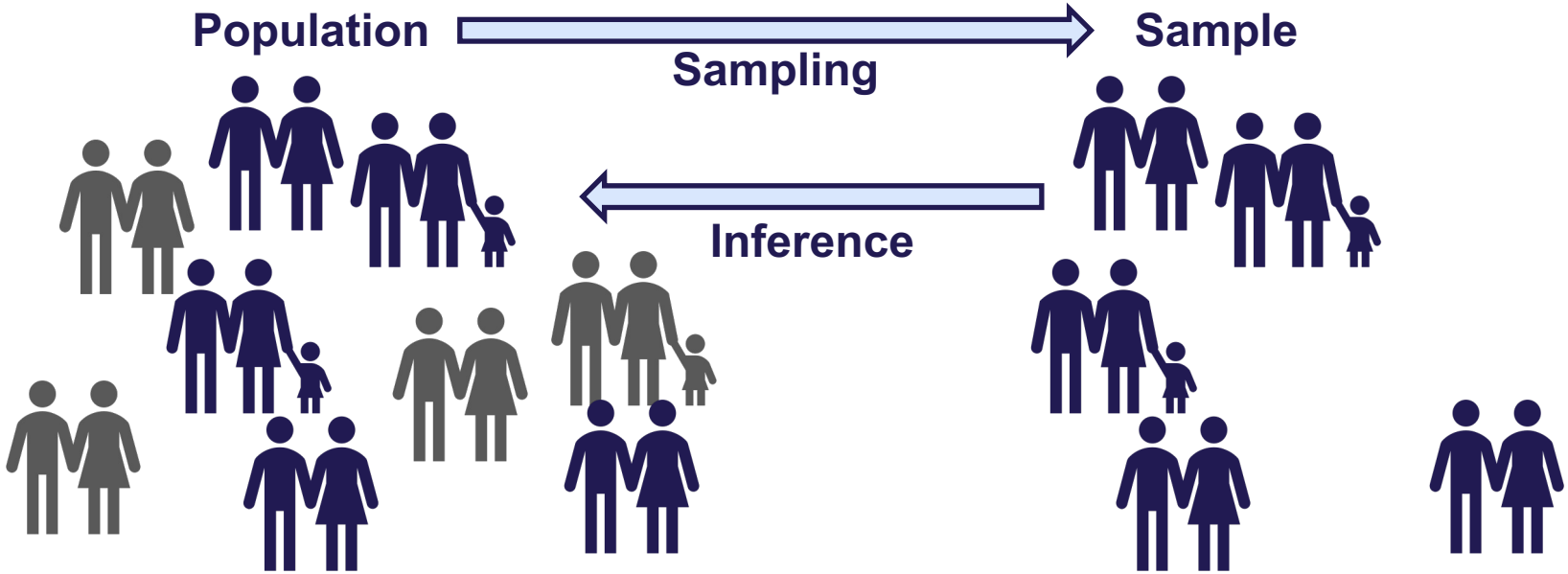
Hard to compute



Sampling

Sampling

If we cannot measure the whole population, we use a smaller sample



Sample space and events

Sample space Ω is the set of possible outcomes of an experiment
Outcomes are individual points $\omega \in \Omega$
Events are subsets of outcomes $A \subseteq \Omega$

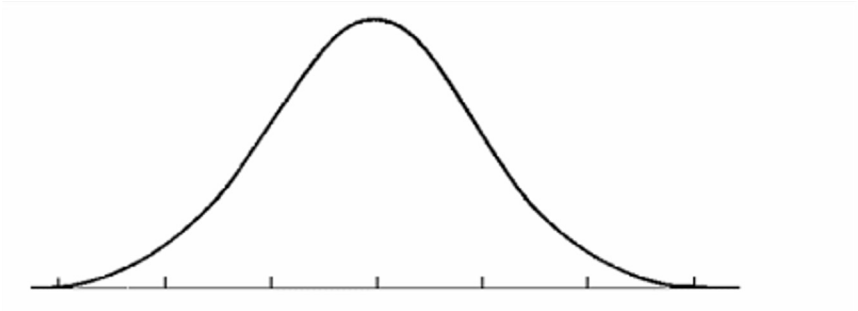
Example

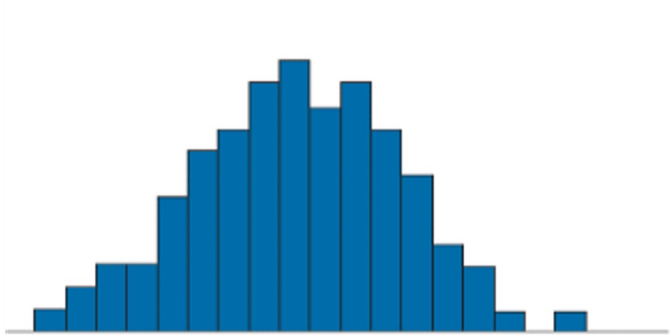
If we toss a coin twice, where the outcome of each toss is either H or T, then

- The possible outcomes are: HH, HT, TH, TT
- The sample space is: $\Omega = \{HH, HT, TH, TT\}$
- The event that the first toss is heads is: $A = \{HH, HT\}$

How do we create a sample?

We randomly draw n values from the population
Each value X_i is a random variable with distribution F
We use the sample to estimate some parameter of F (inference)





Implications and assumptions during sampling

If we randomly draw n values from the population
And each value X_i is a random variable with distribution F
Then a sufficiently large sample will ***look like*** the whole population

What does it mean *looking like*?
The parameters of the population are similar to the statistics
The parameters describe the population and the statistics describe the sample

- Mean
- Variance
- Quantiles

Descriptive statistics

Types of data

Quantitative

Objective measurements about some value: age, height, GPA

Qualitative

Categorical data for classification into groups

Name	Age	Gender	Height	GPA	Degree
Andy Smith	21	M	1.83	94.5%	Electronics
Mark Young	23	M	1.77	93.8%	CS
Alice Brown	19	F	1.79	98.2%	Robotics
Jenny Black	22	F	1.64	94.3%	CS
Luke Scott	18	M	1.92	99.1%	Robotics

Summary representation

Absolute frequency of occurrence

How many samples fall in a category?

Relative frequency of occurrence

Frequency normalized by sample size

Must sum to 1 or 100%

Degree	Frequency	Relative frequency		
CS	31	31/135	0.2296	22.9%
Electronics	23	23/135	0.1703	17.03%
Robotics	44	44/135	0.3259	32.59%
Architecture	37	37/135	0.2741	27.4 %
Total	135			

Histograms

Represent the frequency of occurrence of the outcomes

1. Divide the range of values into bins
2. Count the occurrences
3. Plot

Can be absolute or relative

→ Divide by total no. of values

≈ pdf : divide by bin size

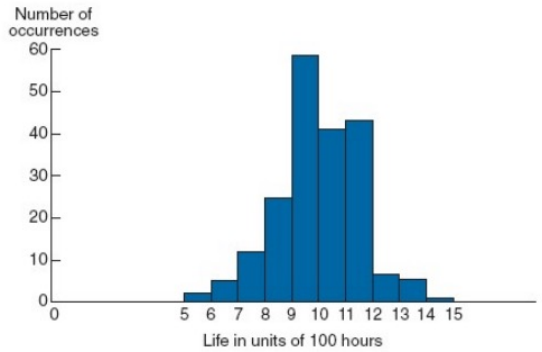
Main challenge

How to choose the bins?

TABLE 2.4 *A Class Frequency Table*

Class Interval	Frequency (Number of Data Values in the Interval)
500–600	2
600–700	5
700–800	12
800–900	25
900–1000	58
1000–1100	41
1100–1200	43
1200–1300	7
1300–1400	6
1400–1500	1

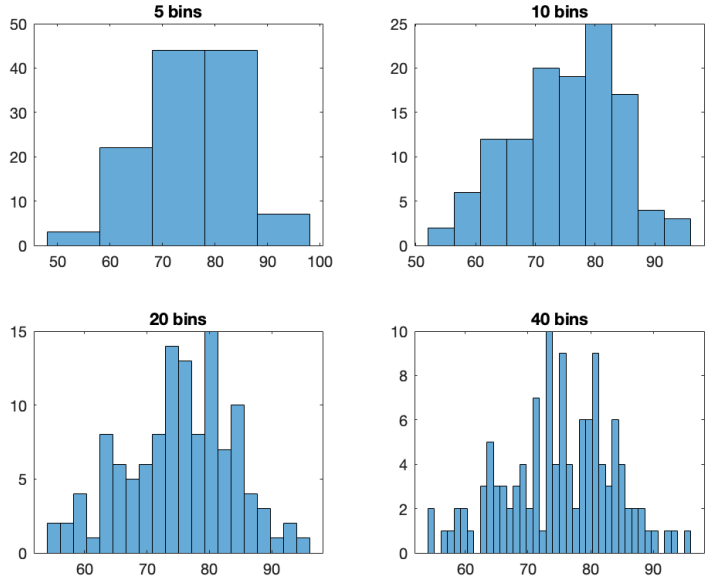
Total X



Example

There is a file called examgrades.mat in the Scripts folder

By running DataDescription.m, we generate the following histograms



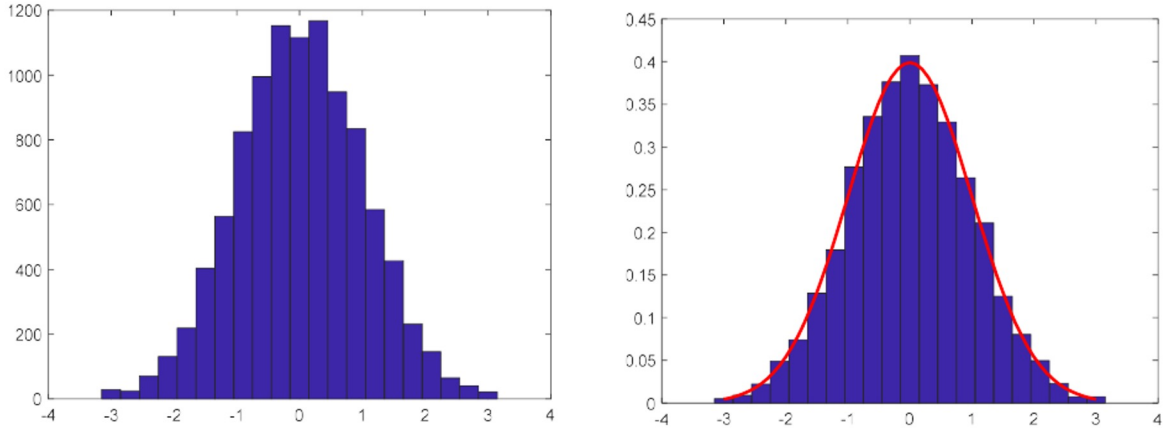
Types of histograms

Absolute frequency

Relative frequency: You get an approximation of a pmf → Σ → CDF

Normalized by bin size: You get an approximation of a pdf

```
Data = randn(10000,1);
```

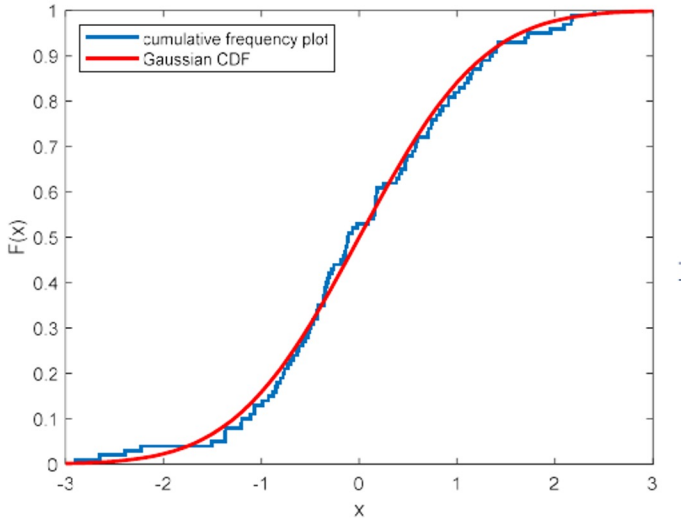


Cumulative frequency plot (ogive)

We know that a CDF shows $P(X \leq x)$: obtained by integrating the PDF/pmf

To obtain the plot:

1. Sort the categories or values (ascending order)
2. The value for a given x is the number of outcomes lesser than x



Sample mean and variance

If $X_1, X_2, \dots, X_n$ are random variables, the **sample mean** is the random variable

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

And the **sample variance** is

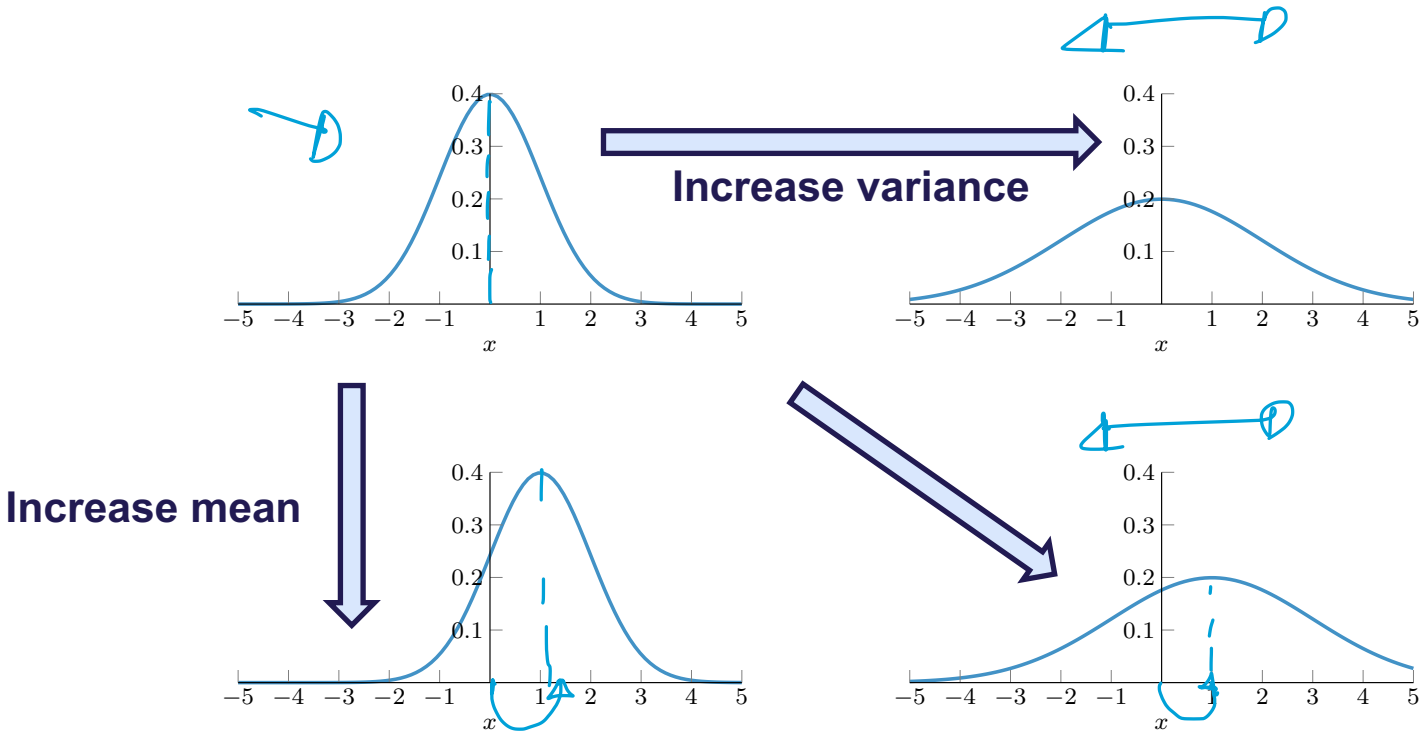
$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X}_n)^2$$

Important

The square root of the variance is the **standard deviation**

Look at the denominator in the formula for the sample variance

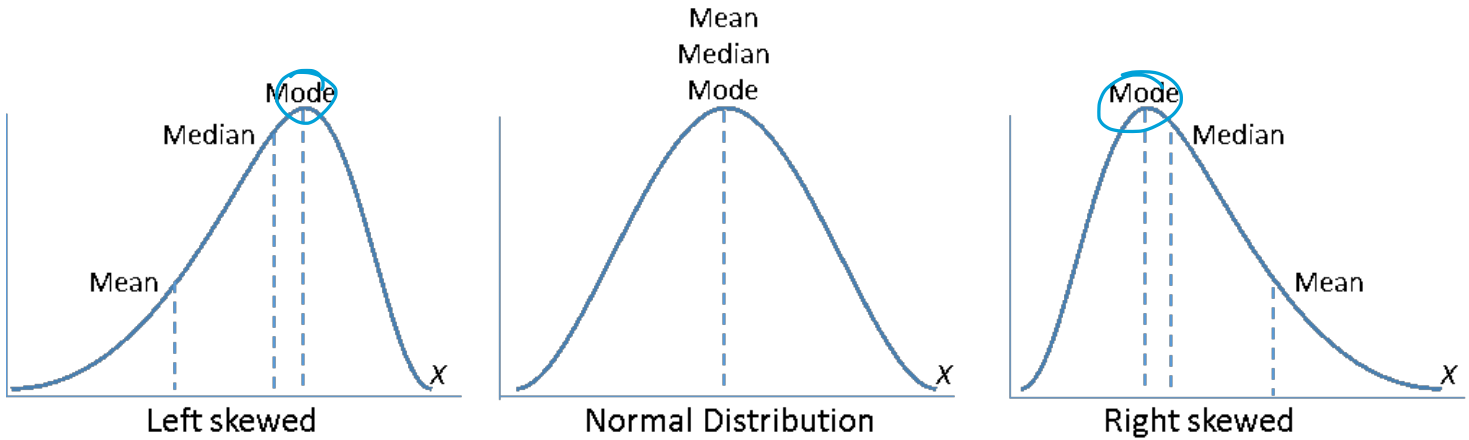
Sample mean and variance visualized



The mean is not always equal to the median

Median: 50% of the outcomes are below and 50% are above

Mode: Category with highest frequency



The law of large numbers

The linearity of the expectation and sample mean

Expectation

$$\mathbb{E}(X+Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

$$\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$$

Sample mean

If $Z_i = X_i + Y_i$, then

$$\bar{Z}_n = \bar{X}_n + \bar{Y}_n$$

If $Z_i = aX_i + b$, then

$$\bar{Z}_n = a\bar{X}_n + b$$

THE LAW OF LARGE NUMBERS

If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables and the sample mean is

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

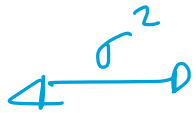
Then, $\bar{X}_n$ **converges in probability** to $\mu = \mathbb{E}(X_i)$ as $n \to \infty$

$$\bar{X}_n \xrightarrow{P} \mu = \mathbb{E}(X_i)$$

So, $\bar{X}_n$ is close to μ with high probability if the sample size is large

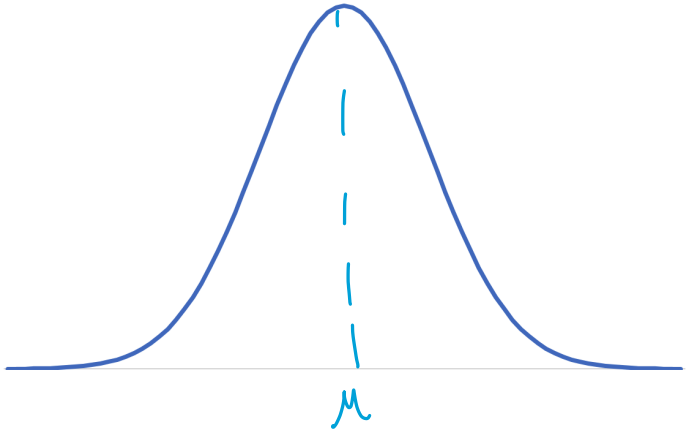
The normal (Gaussian) distribution

The Gaussian distribution



Also called **normal distribution**

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Mean μ, located at the center of the distribution

Variance σ^2 tells us how wide it is

Compact representation is $N(\mu, \sigma^2)$

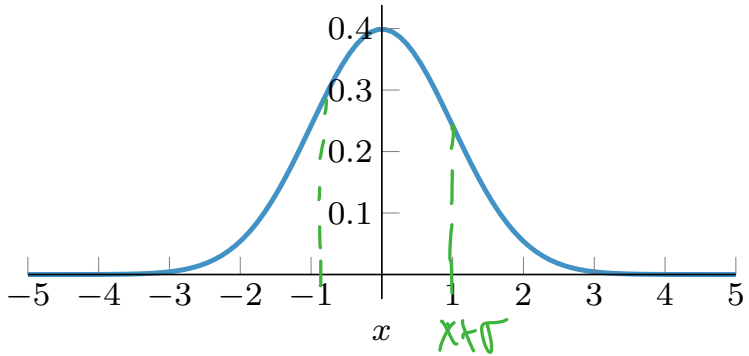
- If X is normal RV we say $X \sim N(\mu, \sigma^2)$

Standard normal distribution $N(0,1)$

The case of a normal distribution with $\mu = 0$ and $\sigma^2 = 1$

Denoted by small greek letter phi

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$



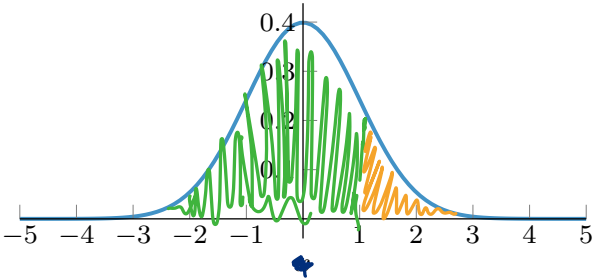
Standard normal distribution and Q-function

Cumulative Distribution Function (CDF) is denoted by $\Phi(x)$ (Phi of x)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \, \mathrm{d}t$$

Q-function

$$Q(x) = \frac{1}{\sqrt{2\pi}} \int_{x}^{\infty} e^{-\frac{t^2}{2}} \, \mathrm{d}t$$



Relation to Q-function

$$\Phi(x) = 1 - Q(x)$$

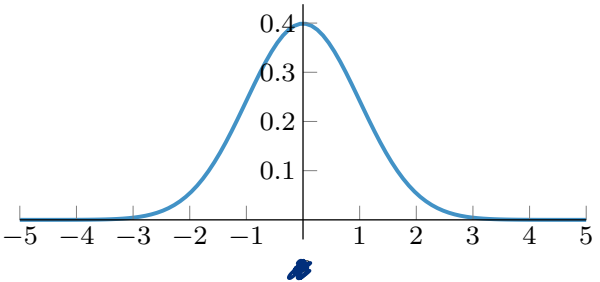
Standard normal distribution and error function

Cumulative Distribution Function (CDF) is denoted by $\Phi(x)$ (Phi of x)

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{t^2}{2}} \mathrm{d}t$$

Error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} \mathrm{d}t$$



Relation to error function

$$\Phi(x) = \frac{1}{2}\left[1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

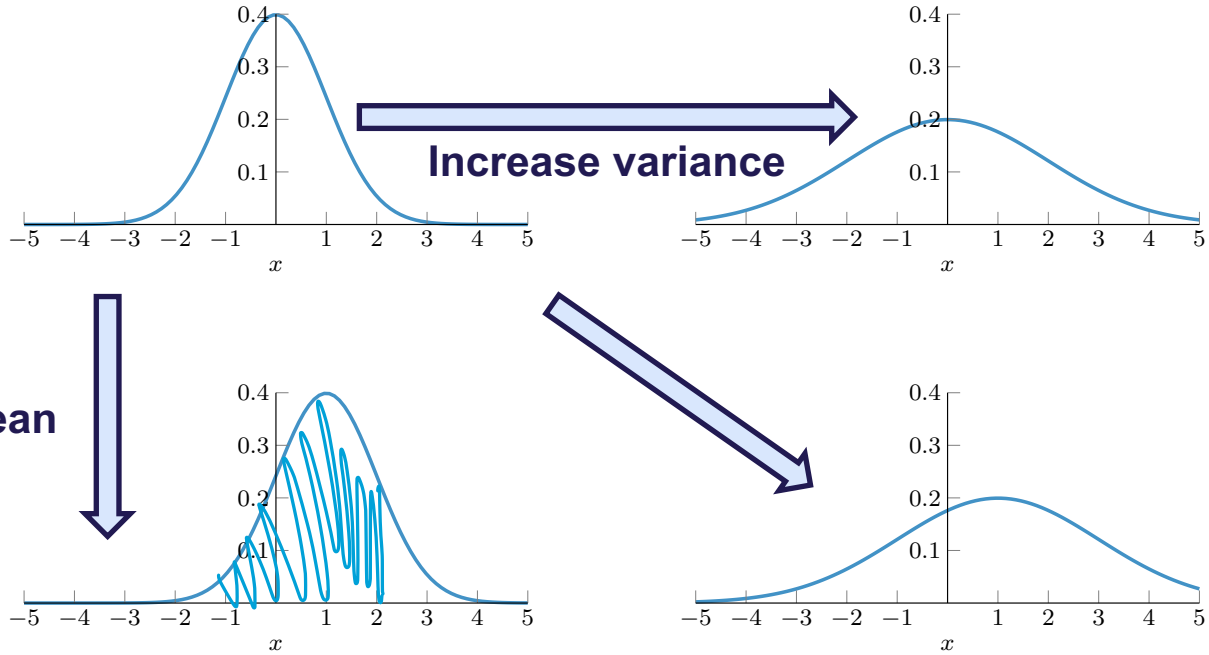
Every normal distribution is a scaled standard normal

PDF

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma}\varphi\left(\frac{x-\mu}{\sigma}\right)$$

CDF

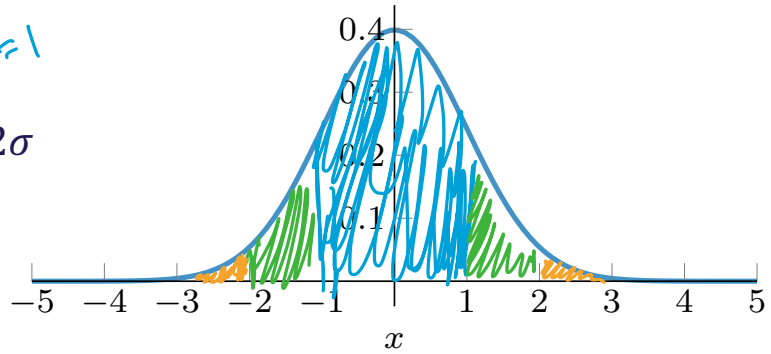
$$F(x;\mu,\sigma^2) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$



Using the standard normal

Probabilities of the outcomes using σ

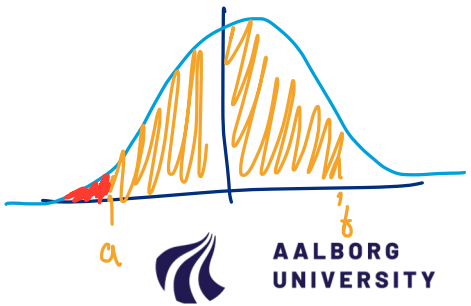
- 68% of outcomes are between $\mu - \sigma$ and $\mu + \sigma$
- 95% of outcomes are between $\mu - 2\sigma$ and $\mu + 2\sigma$
- 99.7% of outcomes between $\mu - 3\sigma$ and $\mu + 3\sigma$



Probability that the outcome of a normal RV $X \sim N(\mu, \sigma^2)$ is between a and b

$$P(a < X < b) = F(b; \mu, \sigma^2) - F(a; \mu, \sigma^2)$$

$$= \Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)$$



Linear operations with the normal distribution

If $X_i \sim N(\mu_i, \sigma_i^2)$ and $i = 1,2, \dots, n$ are independent, then

$$\sum_{i=1}^{n} X_i \sim N\left(\sum_{i=1}^{n} \mu_i, \sum_{i=1}^{n} \sigma_i^2\right)$$

sum of means

sum of variances

If $X \sim N(\mu, \sigma^2)$, then $aX + b$ has mean $a\mu + b$ and standard deviation $|a|\sigma$

Any linear combination of normal RVs X_1 and X_2

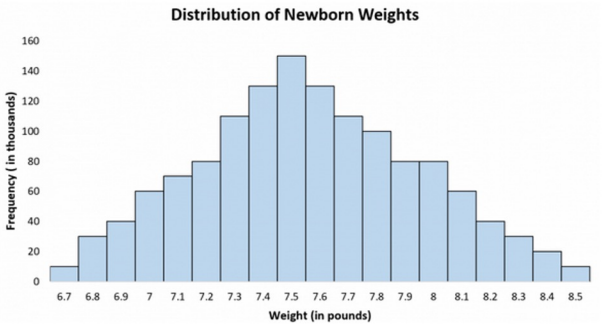
$$W = aX_1 + bX_2$$

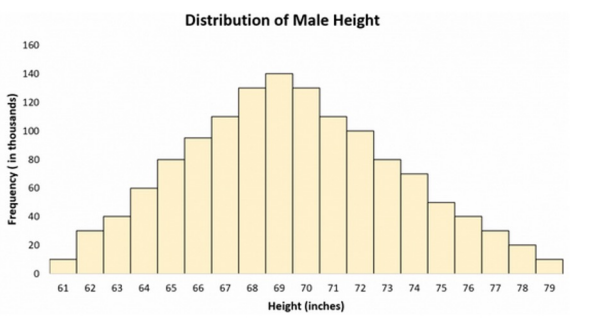
has a normal distribution $W \sim N(a\mu_1 + b\mu_2, a^2\sigma_1^2 + b^2\sigma_2^2)$

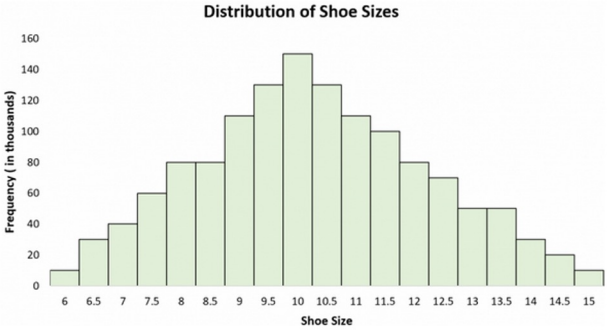
Check normal.py

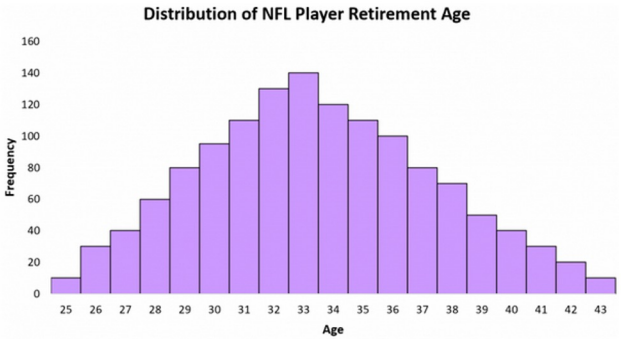
The central limit theorem

Normal distributions are everywhere









Israel Leyva-Mayorga, Introduction to Probability Theory and Statistics, Spring 2023.

AALBORG UNIVERSITY

THE CENTRAL LIMIT THEOREM

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with mean μ and variance σ^2.
Then, if n is large:

$$\sum_{i=1}^{n} X_i = X_1 + X_2 + \cdots + X_n \approx N(n\mu, n\sigma^2)$$

In words

The sum of n i.i.d. random variables is approximately normally distributed with mean $\mu_n = n\mu$ and variance $\sigma_n^2 = n\sigma^2$.

Note: There is no requirement on the type of distribution of $X_1, X_2, \ldots, X_n$

Example: Board game

Variance of $U(a,b)$: $\sigma^2 = \frac{(b-a+1)^2 - 1}{12}$

$3 \times 3.5 \rightarrow$ 3 die

$4 \times 3.5 \rightarrow$ 4 die

$b = 6 \quad a = 1 \Rightarrow \sigma^2 = \frac{35}{12} = 2.91$

A board game requires the player to roll multiple die
The outcome is the sum of values of the die
How does the distribution look like?
One dice: uniform between 1 and 6

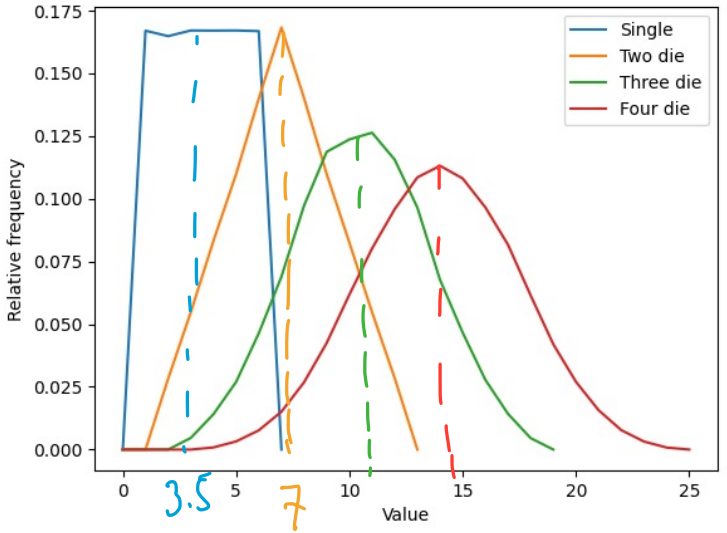
$$X_i \sim U(1,6)$$

$$\mathbb{E}(X_i) = \sum_{k=1}^{6} p_k k = \sum_{k=1}^{6} \frac{1}{6} k = 3.5$$

$2 \times 3.5 \rightarrow$ 2 die

Two, three, four…?
The more die the more it looks normal
Mean $\mu_n = n\mu_1$ Variance $\sigma_n^2 = 2.91n$

$= 3.5n$



Example: Coin toss

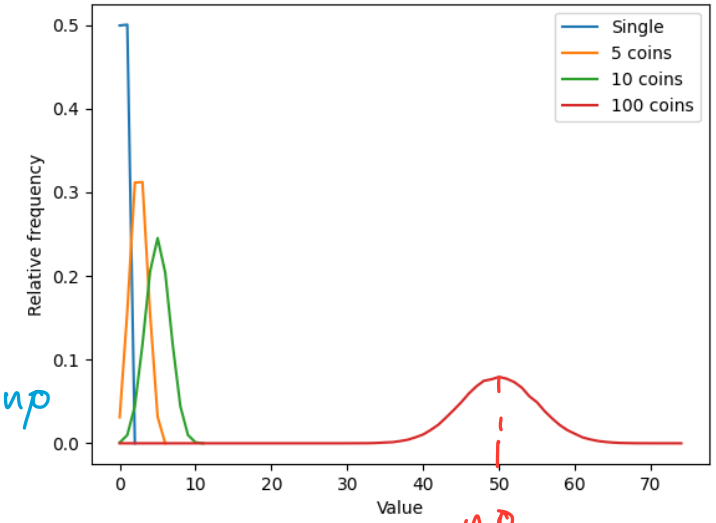
$$P(X=x;n) = \binom{n}{x} p^x (1-p)^{n-x}$$

A coin toss X_i is a Bernoulli RV with probability $p = 0.5$

Outcomes are heads H and tails T

Now we toss the coin n times

The number of heads is a binomial RV H_n



Find the mean of μ_n

$\mu_n = np = 100(0.5) = 50$

- Using the formulas for Binomial distribution
- Using the Central Limit Theorem $\mu_n = n\mu_1 = np$
- Does it match the graph to the right? $\mu_1 = p$

What is the probability of getting $H_{10} = 7$ and $H_{100} = 70$?

Concluding remarks

Lessons learnt

Statistics are everywhere, and it's important to learn them

The basic ways of **describing and visualizing statistical data**:

- Distributions: PDF and pmf
- Cumulative plots: CDF and empirical CDF (ogive)
- Others: box plots, running averages

Sampling is needed and the sample should be representative of the population

Law of large numbers

When the sample is large, we can accurately approximate the mean

The Central Limit Theorem

The sum of a large number of RVs is approximately normal

Useful links and packages

Sources for data:
Our world in data: https://ourworldindata.org/

Matlab
Commands normpdf and normcdf to easily create a normal PDF and CDF

Python
Draw random samples from normal distribution
Numpy: np.random.normal(loc=mu, scale=std, size=(m,n,k))
Scipy: scp.stats.norm.() creates an RV object with the specified parameters

Exercises

Exercise 1

Simulate rolling a fair dice n times.

a) Calculate the sample mean $\bar{X}_i$ and variance S_i^2 for all $i = 1, 2, \dots, n$
b) Plot the sample mean $\bar{X}_i$ and $\bar{X}_i \pm S_i$
c) Plot a normalized histogram with the n outcomes. How many times do you need to roll the dice so the histogram resembles the pmf of a uniform RV?

Help: You can use the code dice_loln.py in the moodle page as a base

Exercise 2

Recall the example of the number of heads after n coin tosses
Also recall that the pmf of a binomial RV is

$$f(n,k,p) = P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},$$

the mean is np and the variance is $np(1-p)$
Calculate the probability of $H_{10} \geq 7$ and $H_{100} \geq 70$:

a) Summing over the pmf as

$$P(H_n \geq k) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i}$$

b) Using the Central Limit Theorem. Are these probabilities close to those in a)?

Exercise 3

The following data represent the lifetime (in hours) of a sample of $n = 40$ transistors:

112, 121, 126, 108, 141, 104, 136, 134, 121, 118
143, 116, 108, 122, 127, 140, 113, 117, 126, 130
134, 120, 131, 133, 118, 125, 151, 147, 137, 140
132, 119, 110, 124, 132, 152, 135, 130, 136, 128

a) Calculate the sample mean $\bar{X}_i$ and variance S_i^2 for all $i = 1, 2, \dots, n$
b) Plot the sample mean $\bar{X}_i, \bar{X}_i \pm S_i$, and $\bar{X}_i \pm 2S_i$
c) Are the data approximately normal?
d) What percentage of the data falls within $\bar{X}_n \pm 2S_n$

Exercise 4

A basketball team will play a 60-game season.
Thirty-two of these games are against class A teams and 28 are against class B teams. The outcomes of the games are independent.
The team will win each game against a class A opponent with probability 0.5, and it will win each game against a class B opponent with probability 0.7.
Let X denote its total number of victories in the season.
a) Is X a Binomial RV?
b) Let X_A and X_B denote, respectively, the number of victories against class A and class B teams. What are the distributions of X_A and X_B?
c) What is the relationship between X_A, X_B, and X?
d) Approximate the probability that the team wins 40 or more games

(normal)