LABORATOIRE HUBERT CURIEN
UMR • CNRS • 5516 • SAINT-ETIENNE
UNIVERSITÉ DE LYON

# From Statistics to Data Mining

## Master 1
## COlour in Science and Industry (COSI)
## Cyber-Physical Social System (CPS2)
## Saint-Étienne, France

Fabrice MUHLENBACH
https://perso.univ-st-etienne.fr/muhlfabr/
e-mail: fabrice.muhlenbach@univ-st-etienne.fr

# Tutorial

- Reminders: Probability and Statistics

- Bayesian method and Bayes rule:
  → make a model of gender prediction from height.

1. collect student data
2. calculate the means and standard deviations
3. carry out a predictive model from the recorded values.

- by setting a discrimination threshold (for the Bayesian predictor of gender from height) set at the middle of the two means of sizes (populations of both sexes), and assuming that the distributions are normal, what probability do we have to be wrong ? (with $\alpha$ and $\beta$ errors!)
- how many errors do you actually get?

# Tutorial

- Reminders: Probability and Statistics

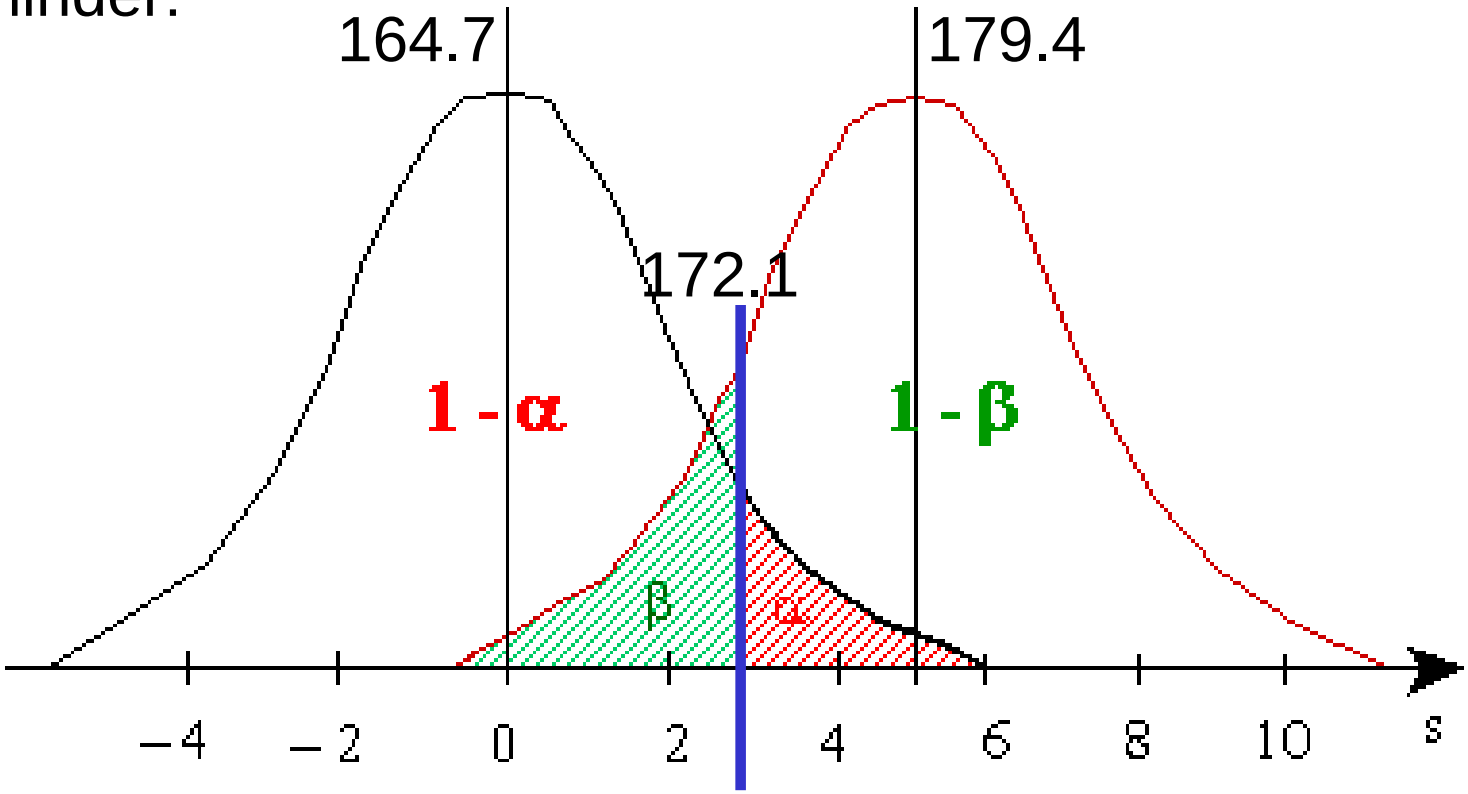
➢ Note: the “true”* values are, for men, an average of 175.6 cm and a standard deviation of 7.2 cm, and for women, an average of 162.5 cm and a standard deviation of 6.2 cm
*: French studies made in 2006

➢ Bonus: https://www.youtube.com/watch?v=gQlqKanUecc

# Tutorial

- Reminders: Probability and Statistics

➢ Reminder:



# Tutorial

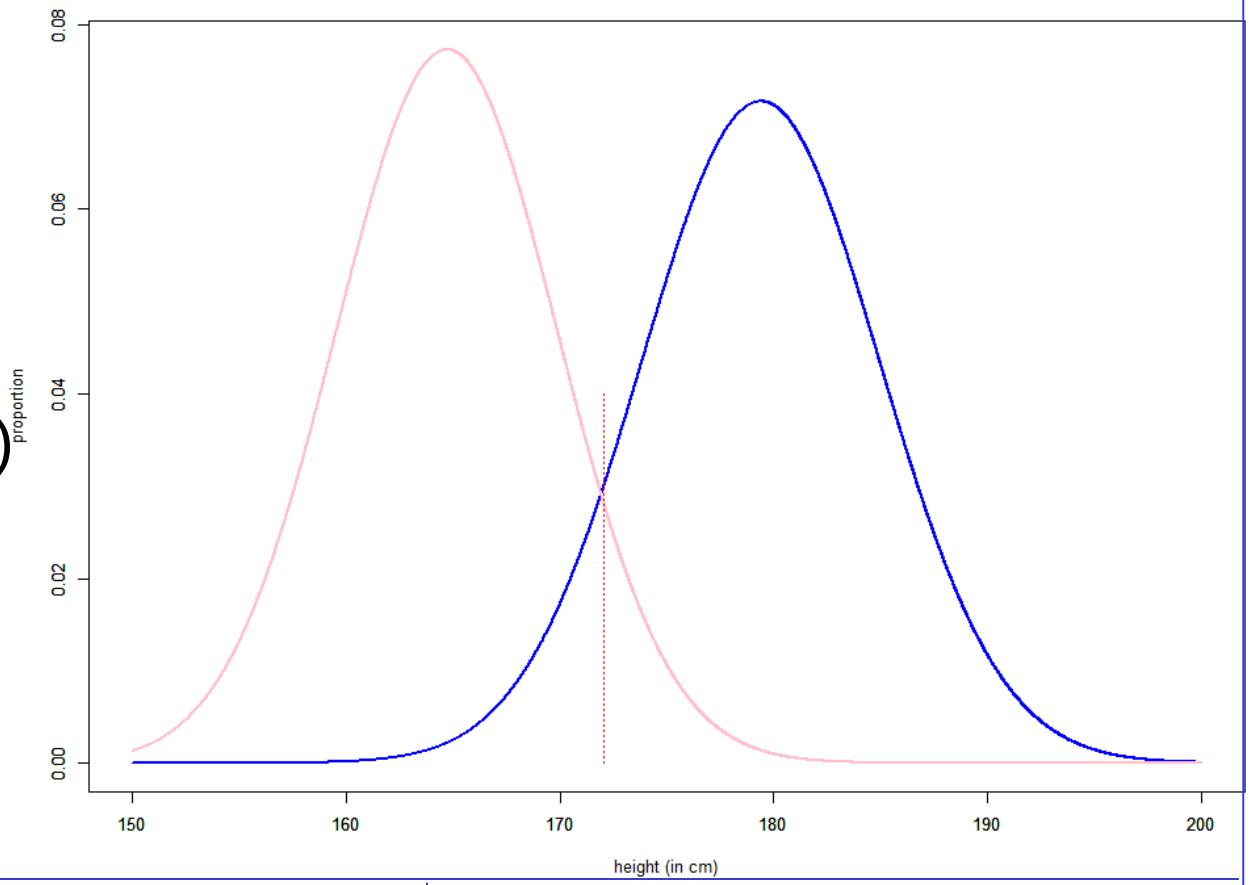
- Reminders: Probability and Statistics

➢ Possible errors:

$p$(woman ∧ tall) + $p$(man ∧ small)

= $p$(woman > 172) + $p$(man < 172)

= 1 - $p$(woman < 172) + $p$(man < 172)



# Tutorial

- Reminders: Probability and Statistics

➢ Computation:

*p*(woman Λ tall) + *p*(man Λ small) with a threshold $\theta$ at 172 cm

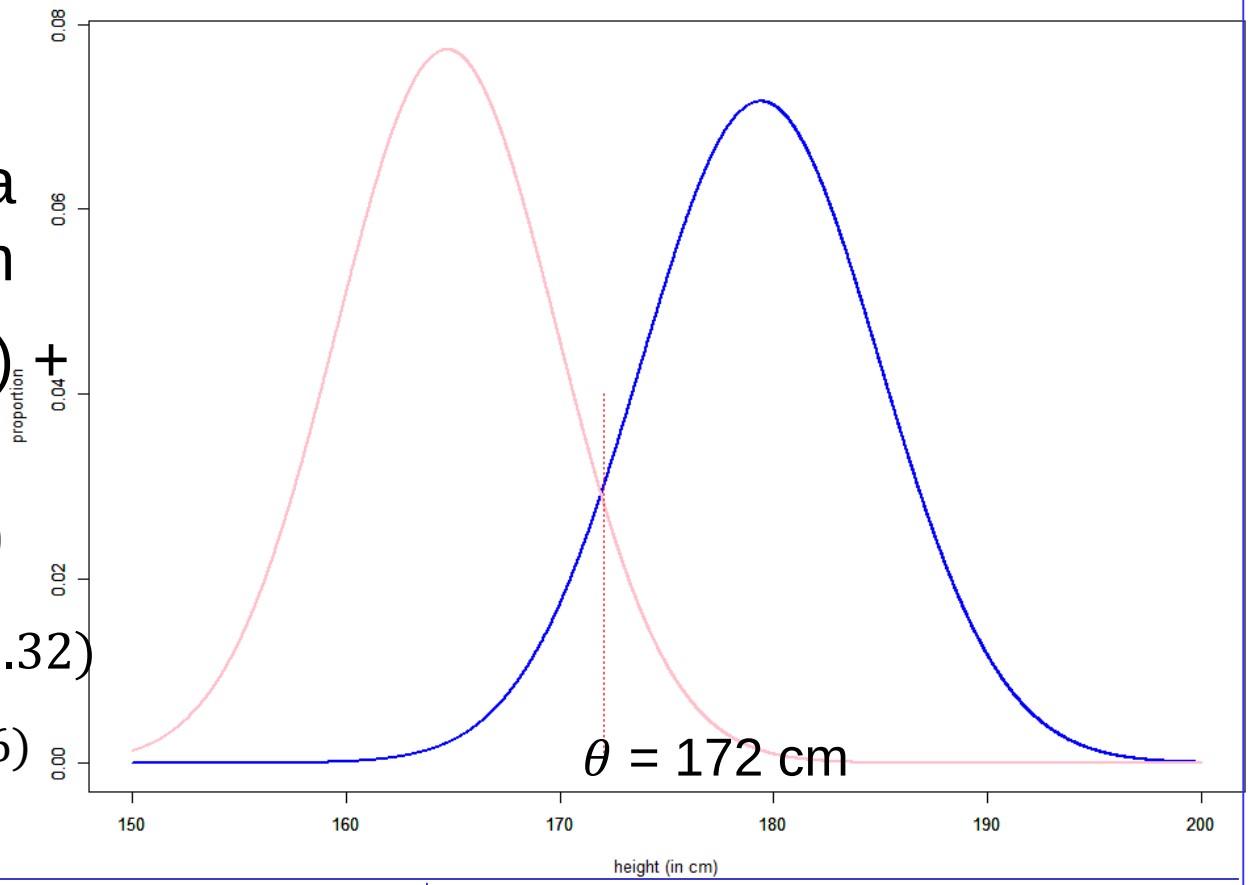
= 1 - *p*(woman < 172) + *p*(man < 172)

$= 1 - p\left(\frac{\theta-\mu_F}{\sigma_F}\right) + p\left(\frac{\theta-\mu_M}{\sigma_M}\right)$

$= p(z > 1.42) + p(z < -1.32)$

$= (1 - 0.9222) + (1 - 0.9066)$

= 17% of error



# Tutorial

- Reminders: Probability and Statistics

➢ Data tabulation
→ from the following data:

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 7500 | 3700 | 2300 | 2200 | 1100 | 530 | 200 | 120 | 30 |

➢ Create the following tables:

1. frequency table
2. cumulative staffing table
3. cumulative frequency table

# Tutorial

- Reminders: Probability and Statistics

➢ Data tabulation

→ 1. frequency table

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $f_i$ | 42.4% | 20.9% | 13.0% | 12.4% | 6.2% | 3.0% | 1.1% | 0.7% | 0.2% |

➢ → 2. cumulative staffing table

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $N_i$ | 7500 | 11200 | 13500 | 15700 | 16800 | 17330 | 17530 | 17650 | 17680 |

➢ → 3. cumulative frequency table

| $x_i$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 |
|---|---|---|---|---|---|---|---|---|---|
| $F_i$ | 42.4% | 63.3% | 76.4% | 88.8% | 95.0% | 98.0% | 99.2% | 99.8% | 100.0% |

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

- we want to study the relationship between the number of hours of work in a course (e.g., from statistics to data mining) and the final grade out of 20 obtained on the examination of this course (to simplify, we will say that there are only 9 possible scores: 0, 2.5, 5, 7.5, 10, 12.5, 15, 17.5 and 20).

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

- The data obtained on the 20 students in the class are as follows:

❑ 2 students did not revise at all and got one 0/20, the other 2.5/20

❑ for the 5 students who studied for 1 hour, there is 1 who obtained 2.5/20, 2 who obtained 5/20, 1 who obtained 7.5/20 and 1 who obtained 10/20

❑ for the 5 students who studied for 2 hours, there is 1 who obtained 5/20, 3 who obtained 10/20 and 1 who scored 12.5/20

❑ for the 6 students who studied for 3 hours, 1 got 7.5/20, 1 got 10/20, 2 got 12.5/20, 1 got 15/20 and 1 got 17.5/20

❑ 2 students revised for 4 hours and got one 17.5/20 and the other 20/20.

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

○ Work to do:

❑ Make an appropriate graphic representation with these data. Be careful, there are several observations with the same values in $X$ and $Y$.
Find an appropriate solution for these data to have a readable and understandable graph.

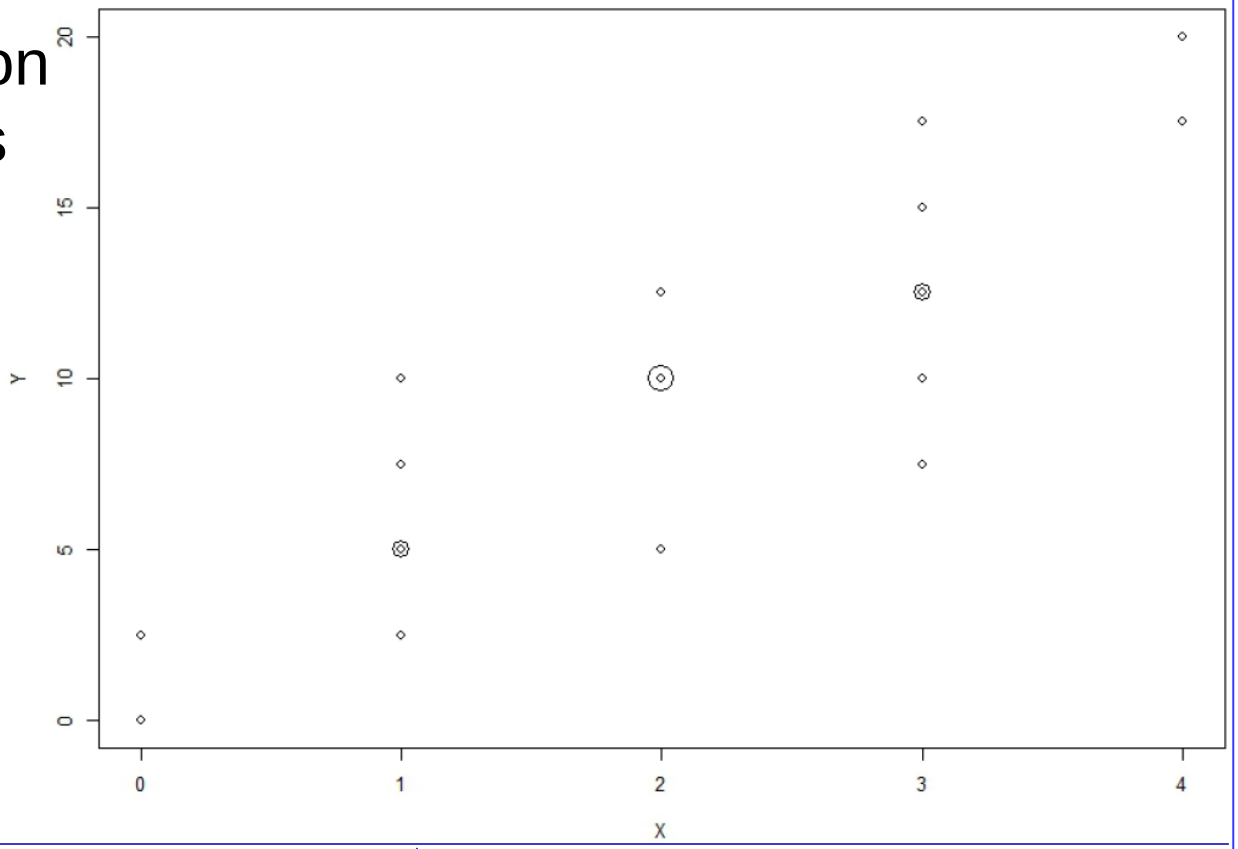
❑ Calculate the correlation coefficient $\rho_{X,Y}$ between $X$ and $Y$.
For this, you can use a contingency table then calculate $p_x$, $p_x.X$, $p_x.X^2$ for $X$, and same for $Y$, and $\sum p_{XY}.X.Y$

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

- Graph representation of the final grade as a function of the revision time
- *X*: revision time (in hours)
- *Y*: final grade (on 20 pts)



# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

○ Contingency table

| $X \backslash Y$ | 0 | 2.5 | 5 | 7.5 | 10 | 12.5 | 15 | 17.5 | 20 |
|---|---|---|---|---|---|---|---|---|---|
| 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 1 | 0 | 1 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 3 | 1 | 0 | 0 | 0 |
| 3 | 0 | 0 | 0 | 1 | 1 | 2 | 1 | 1 | 0 |
| 4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

| $X \backslash Y$ | 0.0 | 2.5 | 5.0 | 7.5 | 10.0 | 12.5 | 15.0 | 17.5 | 20.0 | $p_x$ | $p_x X$ | $p_x X^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 |
| 1 | 0.00 | 0.05 | 0.10 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.25 | 0.25 |
| 2 | 0.00 | 0.00 | 0.05 | 0.00 | 0.15 | 0.05 | 0.00 | 0.00 | 0.00 | 0.25 | 0.50 | 1.00 |
| 3 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.10 | 0.05 | 0.05 | 0.00 | 0.30 | 0.90 | 2.70 |
| 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.10 | 0.40 | 1.60 |
| $p_y$ | 0.05 | 0.10 | 0.15 | 0.10 | 0.25 | 0.15 | 0.05 | 0.10 | 0.05 | 1.00 | $E(X) = 2.05$ | $E(X^2) = 5.55$ |
| $p_y Y$ | 0 | 0.25 | 0.75 | 0.75 | 2.5 | 1.875 | 0.75 | 1.75 | 1 | $E(Y) = 9.625$ | | |
| $p_y Y^2$ | 0 | 0.625 | 3.75 | 5.625 | 25 | 23.4375 | 11.25 | 30.625 | 20 | $E(Y^2) = 120.3125$ | | |
| $\sum_x X.Y.p_{XY}$ | 0 | 0.125 | 1 | 1.5 | 5 | 5 | 2.25 | 6.125 | 4 | $E(XY) = 25$ | | |

➢ $cov(X,Y) = E(XY) - E(X)E(Y) = 25 - 2.05 \times 9.625 = 5.26875$

# Tutorial

- Reminders: Probability and Statistics

➢ Correlation coefficient

| $X\backslash Y$ | 0.0 | 2.5 | 5.0 | 7.5 | 10.0 | 12.5 | 15.0 | 17.5 | 20.0 | $p_x$ | $p_xX$ | $p_xX^2$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.10 | 0.00 | 0.00 |
| 1 | 0.00 | 0.05 | 0.10 | 0.05 | 0.05 | 0.00 | 0.00 | 0.00 | 0.00 | 0.25 | 0.25 | 0.25 |
| 2 | 0.00 | 0.00 | 0.05 | 0.00 | 0.15 | 0.05 | 0.00 | 0.00 | 0.00 | 0.25 | 0.50 | 1.00 |
| 3 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.10 | 0.05 | 0.05 | 0.00 | 0.30 | 0.90 | 2.70 |
| 4 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.05 | 0.05 | 0.10 | 0.40 | 1.60 |
| $p_y$ | 0.05 | 0.10 | 0.15 | 0.10 | 0.25 | 0.15 | 0.05 | 0.10 | 0.05 | 1.00 | $E(X)$ $= 2.05$ | $E(X^2)$ $= 5.55$ |
| $p_yY$ | 0 | 0.25 | 0.75 | 0.75 | 2.5 | 1.875 | 0.75 | 1.75 | 1 | $E(Y)$ $= 9.625$ | | |
| $p_yY^2$ | 0 | 0.625 | 3.75 | 5.625 | 25 | 23.4375 | 11.25 | 30.625 | 20 | $E(Y^2)$ $= 120.3125$ | | |
| $\sum_x X.Y.p_{XY}$ | 0 | 0.125 | 1 | 1.5 | 5 | 5 | 2.25 | 6.125 | 4 | $E(XY)$ $= 25$ | | |

➢ $\sigma_X = E(X^2) - E^2(X) = 5.55 - (2.05)^2 = 1.3475$

➢ $\sigma_Y = E(Y^2) - E^2(Y) = 120.3125 - (9.625)^2 = 27.67188$

➢ $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X.\sigma_Y} = 0.8628273$ → very significant

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- take a coin for coin flipping
- the obverse of a coin is commonly called “heads,” because it often depicts the head of a prominent person, and the reverse “tails”

“heads”  “tails”

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- we will make a set of 10 tosses of the coin
- we will note how many “heads” we will obtain in this set of 10 for each student
- make a graph with these results (→ histogram)
- what is the theoretical probability to obtain 0 “heads,” 1 “heads,” 2 “heads,” ..., 10 “heads”?
- make a graph with the theoretical probabilities
- what are the theoretical mean and standard deviation?
- make a graph with the normal distribution with these values

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ results obtained: nb. of “heads” in 10 tosses of the coin

| $H_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_i$ | 0 | 6 | 16 | 14 | 29 | 29 | 27 | 21 | 5 | 2 | 0 |

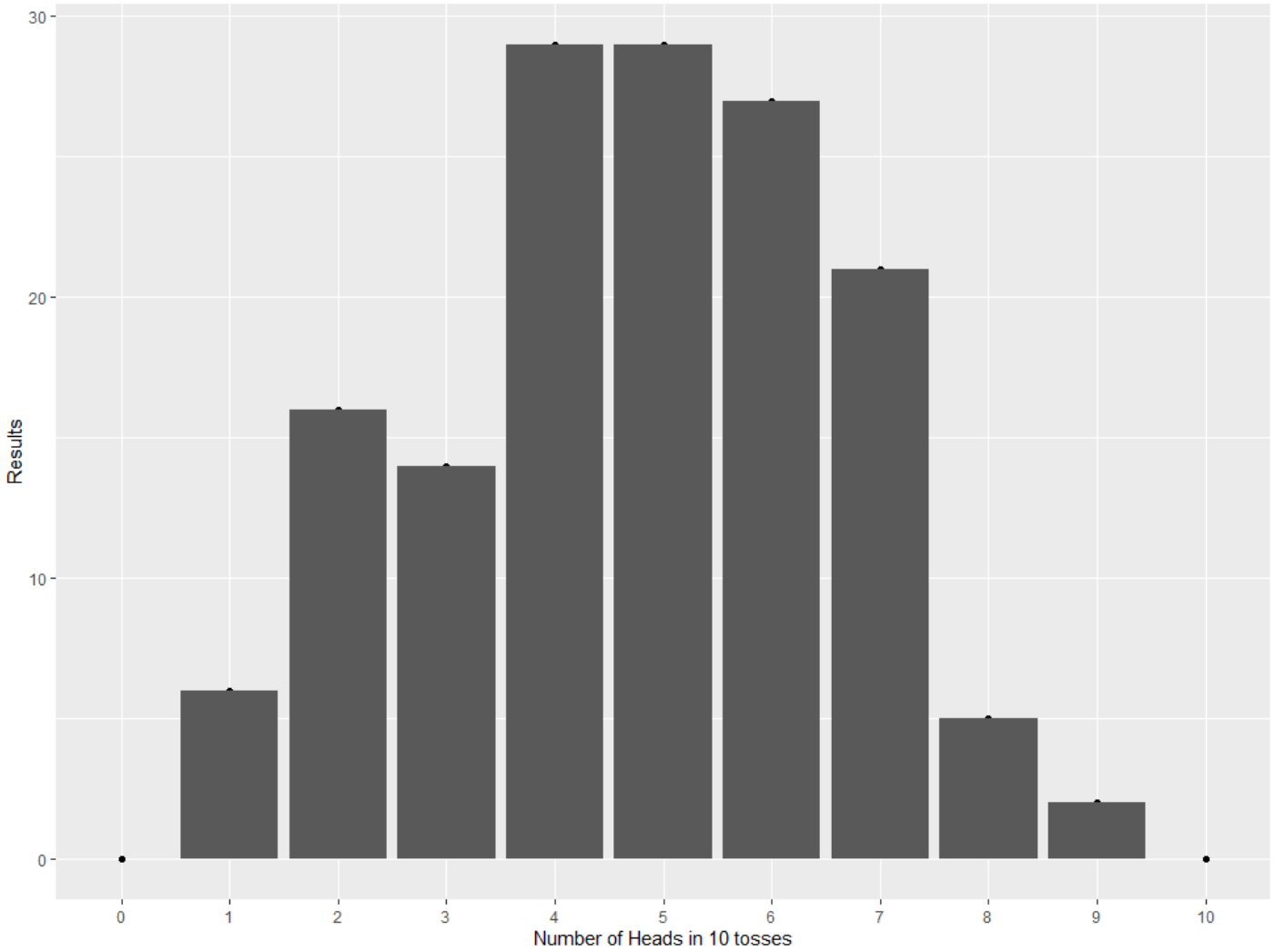
○ total number of results = 149 ≈ 10 repetitions of 10 coin tosses for all 15 students present during the tutorial

○ value of the mean?

○ $\sum_{i=1}^{n} H_i \times \frac{n_i}{n}$

$$= 0 \times \frac{0}{149} + 1 \times \frac{6}{149} + 2 \times \frac{16}{149} + \cdots + 9 \times \frac{2}{149} + 10 \times \frac{0}{149} = 4.75$$

# Tutorial



# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- what is the theoretical probability to obtain 0 “heads,” 1 “heads,” 2 “heads,” ..., 10 “heads”?
- total number of events? (H sides up, 10 tosses)
- H or T, 10 times → $2^{10} = 1024$
- each unique event has $\left(\frac{1}{2}\right)^{10} = 0.0977\%$ chance to occur
- 0 “heads” → T T T T T T T T T T (10 “tails”)

  → $p(0\ \mathrm{H}) = \left(\frac{1}{2}\right)^{10} = 0.0977\%$

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ 1 “heads” →

H T T T T T T T T T

T H T T T T T T T T

...

T T T T T T T T T H

“heads” can occur at the 1st, 2nd, ..., or 10th toss

○ $p(1\ \mathrm{H}) = 10 \times \frac{1}{2} \times \left(\frac{1}{2}\right)^{9} = 10 \times \left(\frac{1}{2}\right)^{10} = 0.9766\%$

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ 2 “heads” →

H H T T T T T T T T

H T H T T T T T T T

...

H T T T T T T T T H

○ → 1st “heads” at the 1st toss

○ → 9 possibilities for the 2nd “heads”

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ 2 “heads” (continued) →

T H H T T T T T T T

T H T H T T T T T T

...

T H T T T T T T T H

○ → 1st “heads” at the 2nd toss

○ → 8 possibilities for the 2nd “heads”

○ for all possibilities: 9 + 8 + 7 + 6 + 5 + 4 + 3 + 2 + 1 = 45

○ $p(2\,\mathrm{H}) = 45 \times \left(\frac{1}{2}\right)^{10} = 4.3945\%$

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- 3 “heads” → selection of items from a collection, such that the order of selection does not matter (unlike permutations)
- → combination of 3 “heads” in a set of 10 tosses
- formula: $\binom{n}{k} = \frac{n!}{k!\,(n-k)!}$
- $\binom{10}{3} = \frac{10!}{3!\times 7!} = 120 \rightarrow p(3\,\mathrm{H}) = 120 \times \left(\frac{1}{2}\right)^{10} = 11.72\%$
- $\binom{10}{4} = \frac{10!}{4!\times 6!} = 210 \rightarrow p(4\,\mathrm{H}) = 210 \times \left(\frac{1}{2}\right)^{10} = 20.51\%$
- $\binom{10}{5} = \frac{10!}{5!\times 5!} = 252 \rightarrow p(5\,\mathrm{H}) = 252 \times \left(\frac{1}{2}\right)^{10} = 24.61\%$

# Tutorial

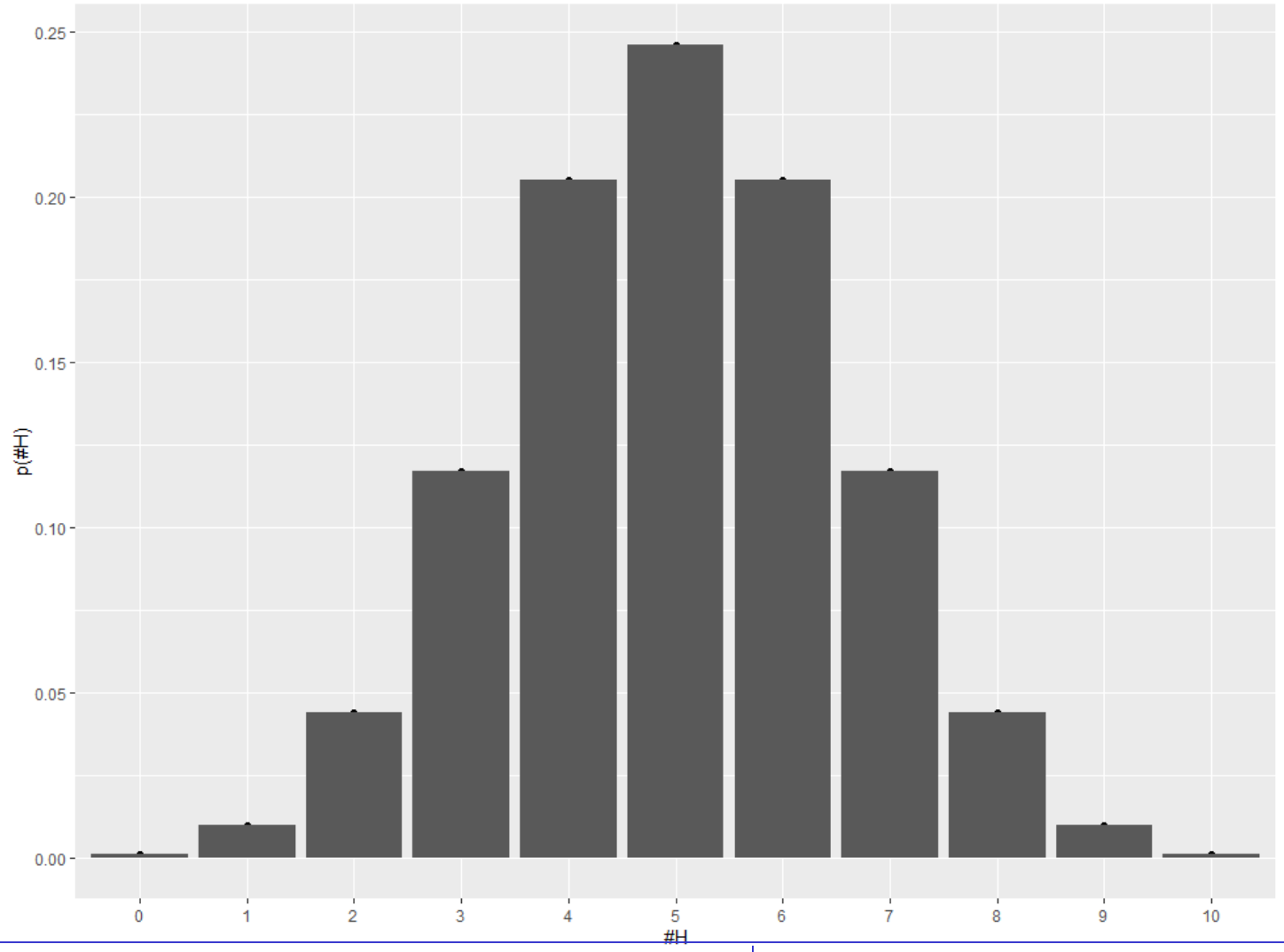
- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ theoretical results obtained:
nb. of “heads” in 10 tosses of the coin (#H)

| $H_i$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $p(H_i)$ % | 0.1 | 1.0 | 4.4 | 11.7 | 20.5 | 24.6 | 20.5 | 11.7 | 4.4 | 1.0 | 0.1 |

# Tutorial



# Tutorial

- Reminders: Probability and Statistics

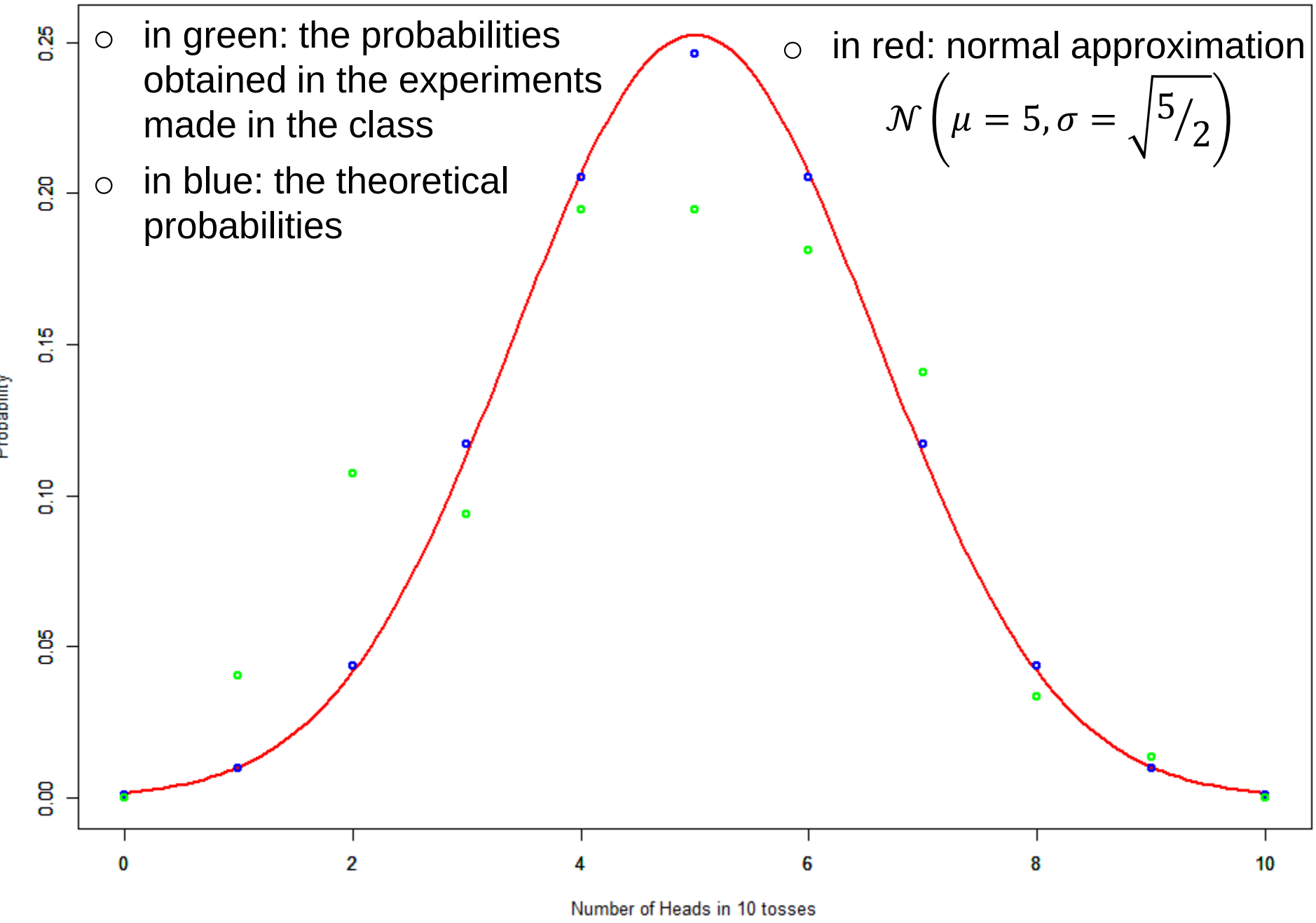
➢ Approximation of a Discrete Variable

- what are the theoretical mean and standard deviation?
- mean $\mu = \sum_{i=0}^{n=10} H_i . p(H_i)$

$$= 0 \times 0.1 + 1 \times 1.0 + 2 \times 4.4 + \cdots = 5$$

- variance $\sigma^2 = \sum_{i=0}^{n=10} (H_i - \mu)^2 . p(H_i) \quad = 5/2$

  → standard deviation $\sigma = 1.58$


in green: the probabilities obtained in the experiments made in the class
in blue: the theoretical probabilities
in red: normal approximation
$\mathcal{N}\left(\mu = 5, \sigma = \sqrt{5/2}\right)$
Probability
0.00
0.05
0.10
0.15
0.20
0.25
0
2
4
6
8
10
Number of Heads in 10 tosses

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

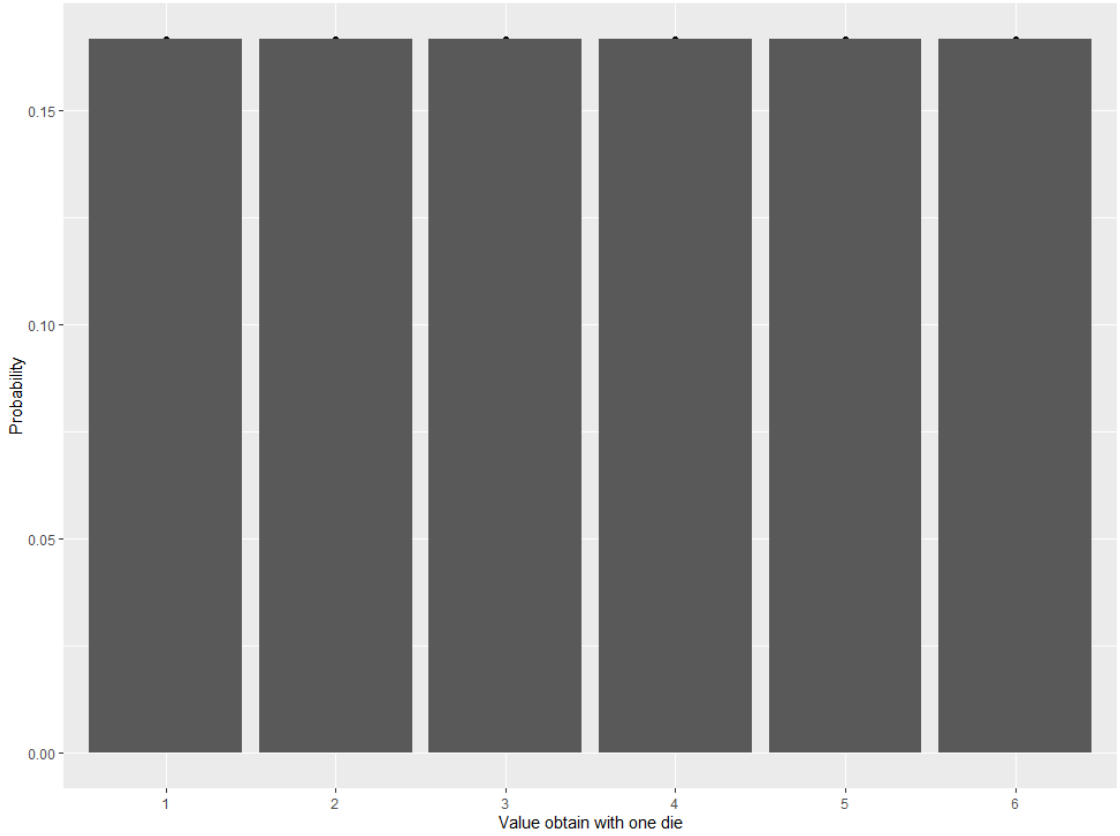
- with 1 die, what is the probability to obtain 1, 2, ..., or 6?
- number of possible events = 6
- number of possible results = 6
- $p(die = 1) = p(die = 2) = p(die = 3) = p(die = 4) = p(die = 5) = p(die = 6) = \frac{1}{6}$

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ with 1 die:



# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- with 2 dice, when we take into account the sum of the dice, what are the possible value? → what are the probabilities?
- number of possible events = 36
  6 possibilities for the 1st die and 6 possibilities for the 2nd die
- number of possible results = 11 (from 2 to 12)
- 2 → {{1,1}}; 3 → {{1,2},{2,1}}; 4 → {{1,3},{2,2},{3,1}};
  5 → {{1,4},{2,3},{3,2},{4,1}}; 6 → {{1,5}, {2,4},{3,3},{4,2},{5,1}};
  7 → {{1,6},{2,5}, {3,4}, {4,3}, {5,2}, {6,1}};
  8 → {{2,6}, {3,5},{4,4},{5,3},{6,2}}; 9 → {{3,6},{4,5},{5,4},{6,3}};
  10 → {{4,6},{5,5},{6,4}}; 11 → {{5,6},{6,5}}; 12 → {{6,6}}

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

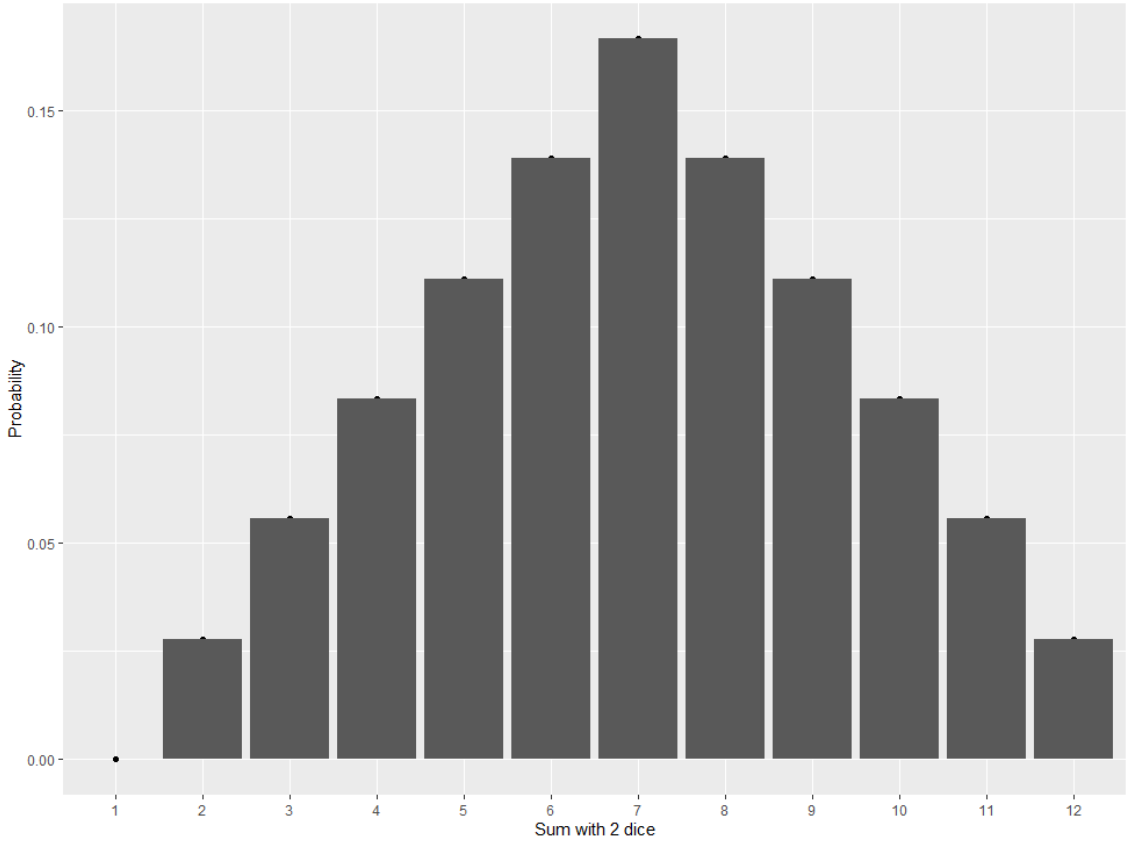
- $p(sum = 2) = p(sum = 12) = \frac{1}{36}$
- $p(sum = 3) = p(sum = 11) = \frac{2}{36}$
- $p(sum = 4) = p(sum = 10) = \frac{3}{36}$
- $p(sum = 5) = p(sum = 9) = \frac{4}{36}$
- $p(sum = 6) = p(sum = 8) = \frac{5}{36}$
- $p(sum = 7) = \frac{6}{36}$

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ with 2 dice:



# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

- with 3 dice:
- number of possible events = 216 (= $6^3$)
- number of possible results = 16 (from 3 to 18)

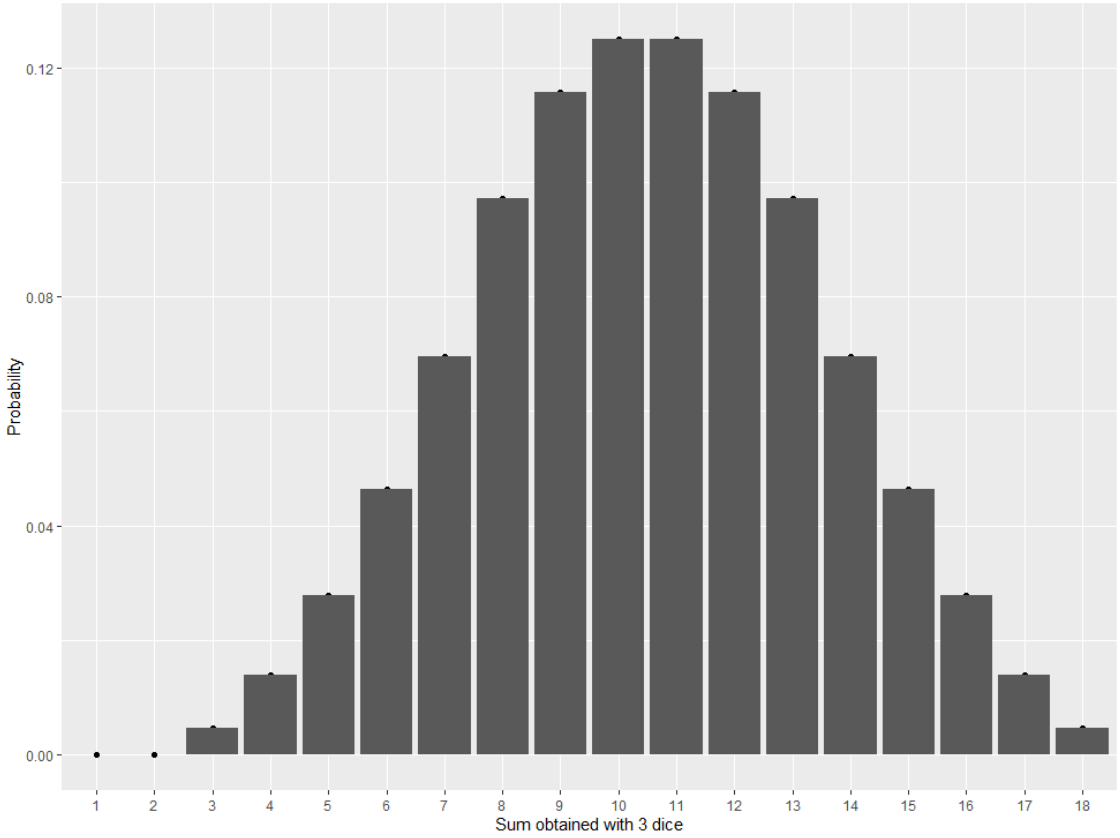
  3 → {{1,1,1}};
  4 → {{1,1,2}, {1,2,1}, {2,1,1}};
  5 → {{1,1,3}, {1,3,1}, {1,2,2}, {2,1,2}, {2,2,1}, {3,1,1}};
  ...
  17 → {{5,6,6}, {6,5,6}, {6,6,5}};
  18 → {{6,6,6}}
  → there are a lot of combinations

# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ with 3 dice:

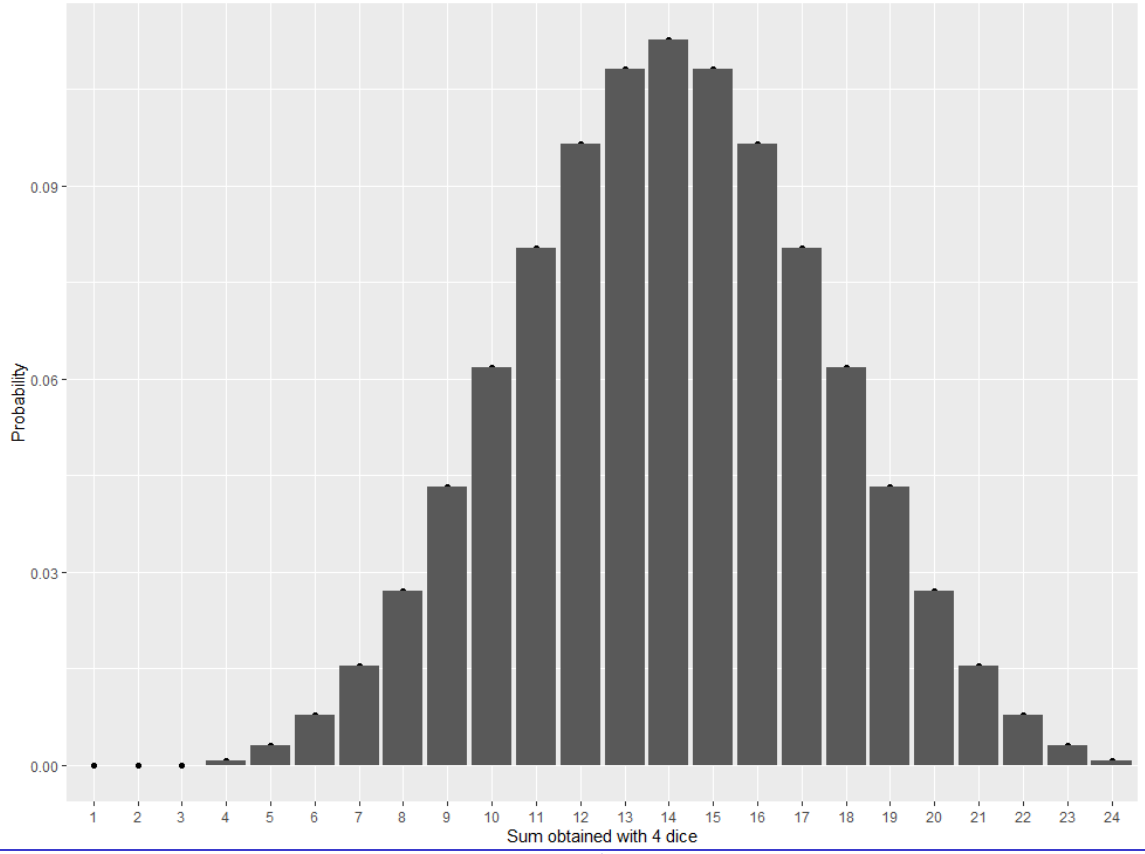


# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ with 4 dice:

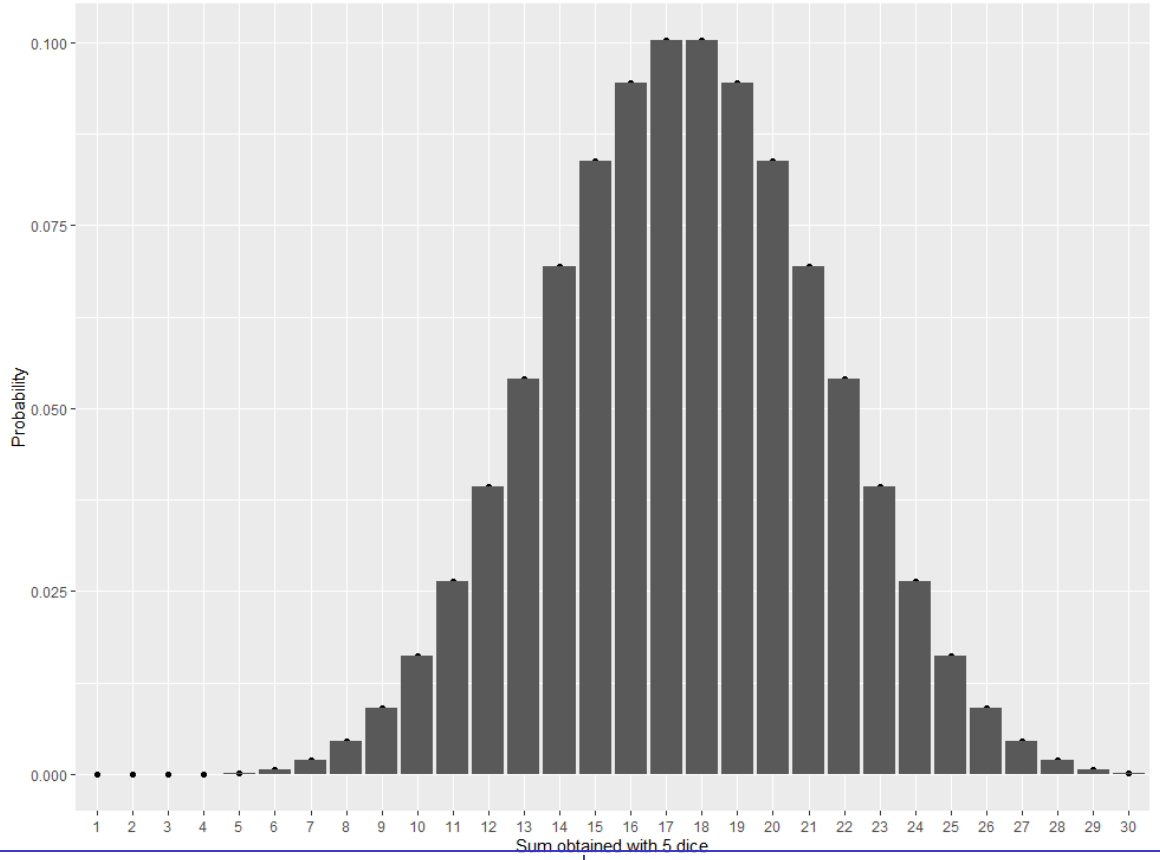


# Tutorial

- Reminders: Probability and Statistics

➢ Approximation of a Discrete Variable

○ with 5 dice:



# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- The motors of household appliances of the *Momo* brand have an average life of $\mu_M$ = 3000 hours with a standard deviation $\sigma_M$= 150 hours.
- Following a change in engine manufacturing, the manufacturer claims that newer engines have a longer average life than older ones
- A sample of $N$ = 50 new engines is tested. We denote by $X_i$ the observed lifetimes. The average (empirical) lifespan of the new engines is calculated: $\bar{X}_M = \frac{1}{50}\sum_{i=1}^{50} X_i = 3040.3$ hours.
- **Question:** Do Momo's new engines bring a statistically significant improvement in household appliance lifespan at a 5% threshold $\alpha$?

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- **Reminder:** when we try to test a mean $\mu$, and when the population follows a normal distribution and the variance $\sigma$ is known, the test statistic $\sqrt{n}\left(\frac{\bar{X}-\mu}{\sigma}\right)$ follows a standard normal distribution $\mathcal{N}(0\,;\,1)$

- **Variant 1:** Same question with a sample of 30 engines instead of 50.
- **Variant 2:** Same question with an average lifespan $\bar{X}_{M'} = 3020$ hours.
- **Variant 3:** Same question with a standard deviation $\sigma'_M$ = 200 hours.
- **Variant 4:** Same question with a threshold $\alpha'$ = 1%.

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- **Reference**: old engines. Lifespan mean: $\mu_M$ = 3000 h, $\sigma_M$ = 150 h.
- **Sample**: 50 new engines. Observed lifespan: $\bar{X}_M$ = 3040.3 h.
- Hypothesis $H_0$: the new process does not change the average lifetime.
- Hypothesis $H_1$: the new process increases the average lifetime.
- How to decide? We use the difference between the observed mean $\bar{X}_M$ and the reference $\mu_M$:

❑ when the difference $\bar{X}_M - \mu_M$ is small, we accept $H_0$

❑ when the difference $\bar{X}_M - \mu_M$ is large enough, we can reject $H_0$ and accept $H_1$.

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- **Question:** is the observed deviation $\bar{X}_M - \mu_M$ = 40.3 h “large enough” to reject $H_0$?
- Intuitively, we understand that the answer will depend on the value of the difference, but also on the size $N$ of the test sample.
- Under $H_0$ hypothesis, the tested engine lifespan is a random variable $X_M$ with a mean $\mu_M$ = 3000 and a standard deviation $\sigma_M$= 150.
- From the Central Limit Theorem, under $H_0$ hypothesis, thanks to the size of the sample $N = 50 > 30$, $\bar{X}_M = \frac{1}{N}\sum_{i=1}^{N=50} X_i \sim \mathcal{N}\left(\mu_M, \frac{\sigma_M^2}{N}\right)$
- $\Rightarrow Z = \frac{\bar{X}_M - \mu_M}{\sigma_M / \sqrt{N}} \sim \mathcal{N}(0,1)$

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- If $H_0$ is true, the standard deviation $Z = \frac{\bar{X}_M - \mu_M}{\sigma_M / \sqrt{N}} \sim \mathcal{N}(0,1)$
- If $H_0$ is true, $p(Z > 1.64) = \int_{1.64}^{+\infty} f_Z(z)dz \cong 0.05 = 5\%$
- The observed standard difference $Z = z_\alpha = \frac{3040.3 - 3000}{150 / \sqrt{50}} = 1.90 > 1.64$
- **Decision**: we reject the null hypothesis $H_0$ (with 5 chances out of 100 to be wrong), therefore we can conclude that the news process increases the lifespan of the household appliance engine.

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- **Variant 1:** Same question with a sample of 30 engines instead of 50.

✓ With 30, the standard difference $Z$ is less important

✓ $Z = z_\alpha = \frac{3040.3-3000}{150/\sqrt{30}} = 1.47 < 1.64$ → we can not reject $H_0$

- **Variant 2:** Same question with an average lifespan $\bar{X}_{M'} = 3020$ hours.

✓ With a smaller lifespan mean, the difference $Z$ is less important too.

✓ $Z = z_\alpha = \frac{3020-3000}{150/\sqrt{50}} = 0.94 < 1.64$ → we can not reject $H_0$

# Tutorial

- Reminders: Probability and Statistics

➢ Statistical hypothesis testing

- **Variant 3:** Same question with a standard deviation $\sigma'_M$ = 200 hours.

✓ With a bigger standard deviation, the standard difference $Z$ is less important

✓ $Z = z_\alpha = \frac{3040.3-3000}{200/\sqrt{50}} = 1.42 < 1.64$ → we can not reject $H_0$

- **Variant 4:** Same question with a threshold $\alpha'$ = 1%.

✓ $Z = z_\alpha = \frac{3040.3-3000}{150/\sqrt{50}} = 1.90 < 2.33$ → we can not reject $H_0$

✓ A value of 1.90 corresponds approximatively to a *p*-value of 3%.