Chapter 2: Probabilities

1 Session 1

Objectives: notion of probability of an event; identify a binomial model; compute probabilities and quantiles from a binomial distribution; understand the law of large numbers: over a large number of experiments, the relative frequency of an event gets close to its theoretical probability.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with the exercises of Session 1 (at least exercises 1, 3).
4. Share and work: discuss your scripts together.
5. Individual work at home: Read the chapter *Probability*, sections *Continuous distributions*.

Exercise 1.1. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is going to be performed on 5 patients. Let X be the random variable equal to the number of successes out of the 5 attempts.

1. Which probability distribution do you propose as a model for X? What are the values? What are the probabilities of the different values? What is their sum?
2. Compute the theoretical mean, variance, standard-deviation, median, first and third quartiles of that distribution. Why are the median and third quartile both equal to 5?
3. What is the probability that the surgery will succeed all 5 times? exactly 3 times? at most 3 times? at least 3 times? from 2 to 4 times?
4. Assign to `X` a simulated sample of size `N=100` of the binomial distribution with parameters 5 and 0.9. Compute the relative frequencies of the different values. Compare with the theoretical probabilities. Repeat (several times) for `N=1e4`, `N=1e6`.
5. Assign to `X` a simulated sample of size `N=1e4`. Plot the ecdf of `X` in blue. Superpose the theoretical cumulative probabilities of the binomial distribution with parameters 5 and 0.9 as red points.

6. Plot as blue dots the cumulative means of `X` (cumulative sums divided by `(1:N)`) against `(1:N)`. Add a horizontal red line marking the theoretical value of the mean of `X`.

7. Plot the cumulative means of `X<=3` against `(1:N)`, as blue dots. Add a horizontal red line marking the theoretical probability.

Exercise 1.2. At an identification session, 6 witnesses are asked to identify a murderer among 4 suspects, including yourself.

1. If each one of the 6 witnesses chooses at random, what are your chances:

   (a) of not being pointed out?

   (b) of being pointed out exactly once?

   (c) of being pointed out twice or more?

2. It turns out that 2 of the 6 witnesses have identified you as the murderer. Referring to 1 (c), do you expect that the judge will think that this may be due to chance?

3. What if 4 of the 6 witnesses have identified you?

Exercise 1.3.

1. Upload `tauber.csv`. Assign column `height` to variable `H`. How many children in the sample are taller than 110 cm? How many have height 110 or less?

2. Plot cumulated frequencies for the event `H>110`, with ordinates in the interval (0,1): plot option `ylim=c(0,1)`.

3. Assign to vector `rH` a random permutation of `H`. Superpose on the same graphics cumulated frequencies for `rH>110`, in blue.

4. Assign to vector `iH` the values of `H`, sorted in increasing order. Superpose on the same graphics cumulated frequencies for `iH>110`, in green. In which interval is the green curve constant?

5. Assign to vector `dH` the values of `H`, sorted in decreasing order. Superpose on the same graphics cumulated frequencies for `dH>110`, in red. In which interval is the red curve constant?

2 Session 2

Objective: compute probabilities and quantiles from normal distributions.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with the exercises of Session 2.
4. Share and work: discuss your scripts together.
5. Individual work at home: read Chapter *Probability*, sections *Fluctuation intervals* and *Limit theorems*.

F. Schena, A. Pattini, S. Mantovanelli (1995) Iron status in athletes involved in endurance and prevalently anaerobic sports. *In C.V. Kies and J.A. Driskel eds., Sports nutrition: minerals and electrolytes, CRC Press, Boca Raton*, p. 69–70.
The article gives means and standard deviations of nutrition and blood parameters, for control populations, and different athlete populations. In all cases, it is assumed that the variable of interest is normally distributed.

Exercise 2.1. The total nutritional intake (in Kcal per day) in the control population has mean 2970 and standard deviation 251. Among runners, the mean is 3350 and the standard deviation 223.

1. Consider a person taken at random in the control population. Would you say that the chances that the intake is smaller than 3000 are larger than 1/2? that it is larger than 3000 are smaller than 1/2? Repeat for runners.
2. What proportion of intakes in the control population are smaller than 2600? larger than 3400? between 2600 and 3400? Repeat for runners.
3. Which lower bound is such that 1% of the control population is below? Which upper bound is such that 1% of runners are above? Use these bounds to represent the two densities on the same plot. Add vertical lines at 2600 and 3400. To which population does the rightmost curve correspond? To which population does the narrower curve correspond? Identify on the graphic the areas corresponding to the probabilities computed in question 2.
4. What intake is such that 5% of the control population is below? above? Repeat for 0.5%, 50%. Repeat for runners.

5. Still using the same lower and upper bounds, represent the two cdf's on the same plot. To which population does the lower curve correspond? To which population does the steeper curve correspond? Where on the graphic could you read the answers to questions 2. and 4.?

6. If one thousand persons were chosen at random in the control population and ranked by increasing intake, how much would the 400-th eat? the 800-th? Repeat for runners.

7. Consider two persons were chosen at random, one from the control population, one from runners. What is the probability distribution for the difference of intakes? Compute the lower bound and the upper bound, such that 1% of the differences are below the lower bound, and 1% above the upper bound. Use these bounds to plot the density of the difference; add green vertical lines at 0 and 400. What would be the chances that the runner eat more than the other? eat over 400 Kcal more per day? Identify the corresponding areas on the graphic.

8. Repeat, replacing runners by cyclists, for whom the mean is 3880 and the standard deviation 450.

9. Repeat, replacing runners by Alpine skiers, for whom the mean is 3524 and the standard deviation 352.

Exercise 2.2. The total carbohydrate intake (in grams per day) in the control population has mean 396 and standard deviation 41. Among runners, the mean is 502 and the standard deviation 36.

1. If a person is taken at random in the control population. Would you say that the chances that the intake is smaller than 400 are larger than 1/2? that it is larger than 500 are smaller than 1/2? Repeat for runners.

2. What proportion of intakes in the control population are smaller than 400? larger than 500? between 400 and 500? Repeat for runners.

3. Which lower bound is such that 1% of the control population is below? Which upper bound is such that 1% of runners are above? Use these bounds to represent the two densities on the same plot. Add vertical lines at 400 and 500. To which population does the rightmost curve correspond? To which population does the narrower curve correspond? Identify on the graphic the areas corresponding to the probabilities computed in question 2.

4. What intake is such that 5% of the control population is below? above? Repeat for 0.5%, 50%. Repeat for runners.

5. Still using the same lower and upper bounds, represent the two cdf's on the same plot. To which population does the lower curve correspond? To which population does the steeper curve correspond? Where on the graphic could you read the answers to questions 2. and 4.?

6. If one thousand persons were chosen at random in the control population and ranked by increasing carbohydrate intake, how much carbohydrate would the 400-th eat? the 800-th? Repeat for runners.

7. If two persons were chosen at random, one from the control population, one from runners: what would be the probability distribution for the difference of intake? Compute the lower bound and the upper bound, such that 1% of the values are below the lower bound, and 1% above the upper bound. Use these bounds to plot the density of the difference; add green vertical lines at 0 and 100. What would be the chances that the runner eat more carbohydrates than the other? eat over 100 grams more per day? Identify the corresponding areas on the graphic.

8. Repeat, replacing runners by cyclists, for whom the mean is 562 and the standard deviation 48.

9. Repeat, replacing runners by Alpine skiers, for whom the mean is 475 and the standard deviation 31.

Exercise 2.3. The red blood cell count (RBC in billion per liter) in the control population has mean 4.96 and standard deviation 0.03. Among runners, the mean is 4.95 and the standard deviation 0.06.

1. If a person is taken at random in the control population. Would you say that the chances that the RBC is smaller than 4.9 are larger than 1/2? that it is larger than 5.0 are smaller than 1/2? Repeat for runners.

2. What proportion of RBC's in the control population are smaller than 4.9? larger than 5? between 4.9 and 5? Repeat for runners.

3. Which lower bound is such that 1% of runners are below. Which upper bound is such that 1% of the control population is above? Use these bounds to represent the two densities on the same plot. Add vertical lines at 4.9 and 5. To which population does the rightmost curve correspond? To which population does the narrower curve correspond? Identify on the graphic the areas corresponding to the probabilities computed in question 2.

4. What RBC is such that 5% of the control population is below? above? Repeat for 0.5%, 50%. Repeat for runners.

5. Still using the same lower and upper bounds, represent the two cdf's on the same plot. To which population does the lower curve correspond? To which population does the steeper curve correspond? Where on the graphic could you read the answers to questions 2. and 4.?

6. If one thousand persons were chosen at random in the control population and ranked by increasing RBC, what would be the RBC of the 400-th? of the 800-th? Repeat for runners.

7. If two persons were chosen at random, one from the control population, and one from runners; what would be the probability distribution for the RBC difference? Compute the lower bound and the upper bound, such that 1% of the values are below the lower bound, and 1% above the upper bound. Use these bounds to plot the density of the difference; add green vertical lines at $0, 0.05$. What would be the chances that the runner has higher RBC than the other, has a RBC over 0.05 grams more per day? Identify the corresponding areas on the graphic.

8. Repeat, replacing runners by cyclists, for whom the mean is 5.01 and the standard deviation 0.04.

9. Repeat, replacing runners by Alpine skiers, for whom the mean is 5.41 and the standard deviation 0.07.

3 Session 3

Objective: compute and interpret one-sided and two-sided fluctuation intervals for normal distributions.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Download function 'normal.fluctuation.R'
3. Share and work: discuss what you understand.
4. Individual work: Practice with the exercises of Session 3.
5. Share and work: discuss your scripts together.

Exercise 3.1. The function `normal.fluctuation` inputs either a probability `p`, or a bound `x` (which can be either a single value, or a vector of two values), a `mean`, a standard-deviation `sd`, and an `alternative` in `"two.sided"`, `"less"`, `"greater"`. It plots a colored area under the density of the normal distribution `dnorm(x,mean,sd)`; it returns an interval `x`, and a probability `p`.

It is known that Dutch 15 year old boys are `mb15=175.2` cm tall on average, with standard-deviation `sdb15=7.9` cm.

1. Download function 'normal.fluctuation.R' and source the function.
2. For each of the following commands: interpret the result, use `pnorm` to compute the same value of `p`.

```
normal.fluctuation(x=160,mean=mb15,sd=sdb15,alternative="greater")
normal.fluctuation(x=185,mean=mb15,sd=sdb15,alternative="less")
normal.fluctuation(x=160,mean=mb15,sd=sdb15,alternative="less")
normal.fluctuation(x=185,mean=mb15,sd=sdb15,alternative="greater")
normal.fluctuation(x=185,mean=mb15,sd=sdb15,alternative="greater")
normal.fluctuation(x=160,mean=mb15,sd=sdb15,alternative="two.sided")
normal.fluctuation(x=185,mean=mb15,sd=sdb15,alternative="two.sided")
normal.fluctuation(x=c(160,185),mean=mb15,sd=sdb15)
```

3. Assign to `S15` a sample of size `n=1e4` of the normal distribution with mean `mb15` and standard deviation `sdb15`. Plot the ecdf of `S14` in blue. Add the cdf of the normal distribution in red. Add green vertical lines at 160 and 185. Identify on the graphic the areas corresponding to the results of the previous question.
4. For each of the following commands: interpret the result, use `qnorm` to compute the same value of `x`.

```
normal.fluctuation(p=0.05,mean=mb15,sd=sdb15,alternative="greater")
normal.fluctuation(p=0.95,mean=mb15,sd=sdb15,alternative="less")
normal.fluctuation(p=0.05,mean=mb15,sd=sdb15,alternative="less")
normal.fluctuation(p=0.95,mean=mb15,sd=sdb15,alternative="greater")
normal.fluctuation(p=0.95,mean=mb15,sd=sdb15,alternative="two.sided")
normal.fluctuation(p=0.025,mean=mb15,sd=sdb15,alternative="less")
normal.fluctuation(p=0.025,mean=mb15,sd=sdb15,alternative="greater")
```

5. Use the sample of question 2 to check the results, like in question 3.

6. Compute two-sided fluctuation intervals for the normal distribution with mean `mb15` and standard deviation `sdb15`, with levels 0.90, 0.95, 0.99. Why does the width of the interval increase with the level?

7. Repeat for 15 year old Dutch girls, for which the mean height is `mg=166.9` cm, with standard deviation `sdg=6.8` cm.

Exercise 3.2. Recall from the previous lab session that the total nutritional intake (in Kcal per day) in the control population has mean 2970 and standard deviation 251. Use `normal.fluctuation` to compute the following quantities. Check the results using `pnorm` or `qnorm`.

1. What is the proportion of control persons with intake larger than 3500? smaller than 3200? smaller than 2800? between 2800 and 3400?

2. What is the intake value such that 90% of the population is below? above? Repeat with 95%.

3. Compute two-sided fluctuation intervals with levels 0.90, 0.95, 0.99.

4. Repeat with runners, for whom the mean is 3350 and the standard deviation 223.

5. Repeat with cyclists, for whom the mean is 3880 and the standard deviation 450.

6. Repeat with Alpine skiers, for whom the mean is 3524 and the standard deviation 352.

7. Of which athletes would you say they eat significantly more than the control population?

Exercise 3.3. Recall from the previous lab session that the red blood cell count (RBC in billion per litre) in the control population has mean 4.96 and standard deviation 0.03. Use `normal.fluctuation` to compute the following quantities. Check the results using `pnorm` or `qnorm`.

1. What is the proportion of control persons with RBC larger than 4.9? smaller than 5? between 4.95 and 5.02?

2. What is the value of RBC such that 90% of the population is below? above? Repeat with 95%.
3. Compute two-sided fluctuation intervals with levels 0.90, 0.95, 0.99.
4. Repeat with runners, for whom the mean is 4.95 and the standard deviation 0.06.
5. Repeat with cyclists, for whom the mean is 5.01 and the standard deviation 0.04.
6. Repeat with Alpine skiers, for whom the mean is 5.41 and the standard deviation 0.07.
7. Of which athletes would you say their RBC is larger than the control population?

4 Session 4

Objectives: understand the meaning of q-q plots. Understand normal approximations. Compare probabilities and quantiles computed from a binomial distribution, and from its normal approximation.

Steps of the session:

1. Individual work: Practice with the instructions given on the web site.
2. Share and work: discuss what you understand.
3. Individual work: Practice with the exercises of Session 4 (at least exercises 1 and 2).
4. Share and work: discuss your scripts together.
5. Individual work at home: Read the chapter *Estimation and confidence intervals*.

Exercise 4.1.

1. It is known that Dutch 15 year old boys are `mb=175.2` cm tall on average, with standard-deviation `sdb=7.9` cm. Draw a sample `X` of size `N=10^4` from the normal distribution with mean `mb` and standard deviation `sdb`. Display its first 100 values. Assign to `Xs` the values of `X`, sorted in increasing order (function `sort`). Display the first 100 values of `Xs`.
   (a) Plot the points with abscissas `Xs`, ordinates `(1:N)/N`. Which function did you represent?
   (b) Plot the points with abscissas `(1:N)/N`, ordinates `Xs`. Which function did you represent?
   (c) Plot the points with abscissas `qnorm(((1:N)-0.5)/N)`, ordinates `Xs`. Can you explain why the points are close to a straight line? Which straight line are the points close to? Add that straight line to the plot.
   (d) Run `qqnorm(X)`. Add the same straight line. Is there any difference with the previous plot?
2. Draw a sample `X` of size `N=10^4` of the Student distribution with parameter `df=5`. Run `qqnorm(X)`. Repeat for `df=10`, then `df=50`. What do you observe? Which straight line are the points close to in the last case? Add that straight line to the plot.
3. Recall that the chi-squared distribution with parameter `df` has mean `df`, standard deviation `sqrt(2*df)`. Draw a sample `X` of size `N=10^4` of the chi-squared distribution with parameter `df=5`. Run `qqnorm(X)`. Repeat for `df=50`, then `df=500`. What do you observe? Which straight line should the points be close to? Add that straight line to the plot.

4. Recall that the binomial distribution with parameters `n` and `p` has mean `n*p`, standard deviation `sqrt(n*p*(1-p))`. Draw a sample `X` of size `N=10^4` of the binomial distribution with parameters `n=50`, `p=0.5`. Run `qqnorm(X)`. Repeat for `n=500`, then `n=5000`. What do you observe? Which straight line should the points be close to? Add that straight line to the plot.

Exercise 4.2. From past experience, it is known that a certain surgery has a 90% chance to succeed. This surgery is performed by a certain clinic 400 times each year. Let N be the number of successes next year.

1. What is the exact model for N? What is the normal approximation?
2. What are the probabilities for the clinic to perform successfully the surgery at least 345 times, in the exact and approximate models.
3. The insurance accepts to cover a certain number of failed surgeries; that number has only a 1% chance to be exceeded. What number is it, in the exact and approximate models?

Exercise 4.3. Among people old enough to receive an injection against the flu, 40% of them ask for it. In a population of 150000 persons old enough to receive the injection, let N be the number of those that will ask for it.

1. What is the exact model for N? What is the normal approximation?
2. If 60500 syringes are prepared, what is the probability that these will not suffice, in the exact and approximate models?
3. Find the number of syringes that should be prepared to ensure that there will be enough with 90% probability at least, in the exact and approximate models.

5 Session 5

Objectives: compute probability characteristics on your dataset.

Steps of the session:

1. Team work: Compute fluctuation intervals, probabilities, quantiles and qqplot on your project dataset.
2. Team work: Write the report corresponding to your results
3. Individual work at home: Read the chapter *Estimation and confidence intervals*.