Exercise Set 4

Goals

1) To explore the law of big numbers and appreciate this effect on the precision of a measurement.
2) To understand how it can be used in risk assessment (exercise on the car insurance).

1 Expectation values and variances

Let events/subset X_1, X_2, ..., X_n follow the normal law $\mathcal{N}(\mu, \sigma^2)$ and all be *independent* with respect to each other. $\mathbb{E}$ denotes the expectation value. Calculate:

a) $\mathbb{E}(10X_1 + 2X_2)$
b) $\mathbb{E}((X_1 + X_2 + ... + X_n)/n)$
c) $\mathrm{Var}(10X_1 + 2X_2)$
d) $\mathrm{Var}((X_1 + X_2 + ... + X_n)/n)$

2 The influence of replicates on the precision of a measure: the average chemical composition of the interstellar medium

According to the Big Bang theory, the universe should be composed of about 75% of Hydrogen. Two competing Big Bang models predict two slightly different compositions, the first one 74% and the second one 77%.

One single measurement of the interstellar composition is precise to ±1%. Each such satellite-based measurement is extremely expensive and time consuming. How many measurements are sufficient to distinguish between the two models with 99.99% of confidence?

a) Let X be this random experiment modelled by a normal law. If the universe follows theory A, the measurement results should be distributed as $\mathcal{N}(74, 1)$. Within that model, what is the probability of measuring 77% or higher?
b) Inversely, if nature is following $\mathcal{N}(77, 1)$ what is the probability of measuring 74% or lower?
c) Compute for both cases, the corresponding 99.99% confidence intervals.
d) If the law is $\mathcal{N}(74, 1)$ for one single measurement, what is the law for an average of ten measurements, $\bar{X}_{10} = \frac{1}{10}\sum_{i=1}^{10} X_i$?
e) Let us assume we measure one single measurement at 75%. What is the 99.99% confidence interval? Are 77% or 74% inside?

   Let us assume that is was not the result of one single measurement, but ten repeated ones, $\bar{X}_{10}$. How does the interval reduce? How about for 1000 averaged measurements?

f) Let us assume the average over n measurements results in 75%. How many replicates (n) would be sufficient so that 77% would just lie outside the 99.99% confidence interval, thus supporting the 74% theory?

3 Bernoulli law, Normal law and Car insurance

Table 1 shows a list of different car accidents in Switzerland and their average costs for the insurance. The total operating budget of all insurance firms (paying all employees, renting the building, etc. but not the payout of insurance cases) is 1.7 billion CHF, and the total number of insured cars in Switzerland is 4.4 million.

a) How much does the average client cost a company? Model this as a Bernoulli random experiment assuming that a client has either an accident in a given year, or not (ignore multiple accidents). Under this assumption, estimate the parameter p.

b) We now model the distribution by a Gaussian of center p and of variance $p(1-p)/n$, following the Binomial law. How much should a client pay such that an insurance with $10'000$ clients is profitable with a probability of 95%?

c) How would b) change with $1'000'000$ clients?

d) The average yearly insurance cost in Switzerland is $1'100$ CHF. What profit do car insurance companies make?

What is the profit per employees, if there are $10'000$ employees for car insurance?

219 dead	$300'000$ CHF
3654 seriously injured	$100'000$ CHF
$17'759$ lightly injured	$5'000$ CHF

Table 1: Yearly accidents on the swiss roads and respective average costs.

4 Molecular Beam Epitaxy (MBE), Normal law and precision

A lab has developed an MBE technique to control the thickness of a film on a silicon wafer (see Dataset 1).

a) Compute the median $\tilde{x}$, mean $\bar{x}$ and standard deviation s of this sample dataset 1. Are there any hints for an asymmetric distribution?

b) Draw a histogram (choose an adequate bin size): is there only one population or more? outliers?

c) What fraction of the data lies within the intervals $[\bar{x} \pm s]$ and $[\bar{x} \pm 2s]$?

d) Model the data by a Gaussian distribution using the experimental values for mean and deviation, assuming $\sigma = s$ and $\mu = \bar{x}$. Compare the fractions computed in c with the ones of this Gaussian model, $[\mu \pm \sigma]$ and $[\mu \pm 2\sigma]$. What are the differences to c)? Why? Can the distribution be represented by a Gaussian?

e) It turns out that the results were achieved by by two different technicians. One of the two repeats the experiment another time (Dataset 2).

f) Recompute the questions a) to d) with the second set and discuss whether the quality has increased or decreased?

g) Intel will only start a partnership with the lab if 95% of the chips lie within the interval [1.45, 1.55]. Assuming that you can infer the correct Normal/Gaussian distribution expected for future production based on this data: Can the lab certify this requirement for its chips? What is problematic about this inference?

h) Compute the 5% and the 95% quantiles? How does this compare against the experimental bounds that Intel required?

Dataset 1 (film thickness [nm]):

1.5	1.71	1.62	1.81	1.27	1.91	1.05	1.15	1.22	1.25	2.00	1.95
1.82	1.84	1.81	1.42	1.37	1.86	1.29	1.85	1.26	1.87	1.84	1.83

Dataset 2 (film thickness [nm]):

1.33	1.71	1.76	1.44	1.45	1.46	1.46	1.62	1.63	1.65	1.67	1.67	1.51
1.51	1.52	1. 58	1.52	1.57	1.57	1.56	1.55	1.56	1.54	1.53	1.53	