MAP531 : Statistics
PC2 – Statistical modeling

1 To warm up

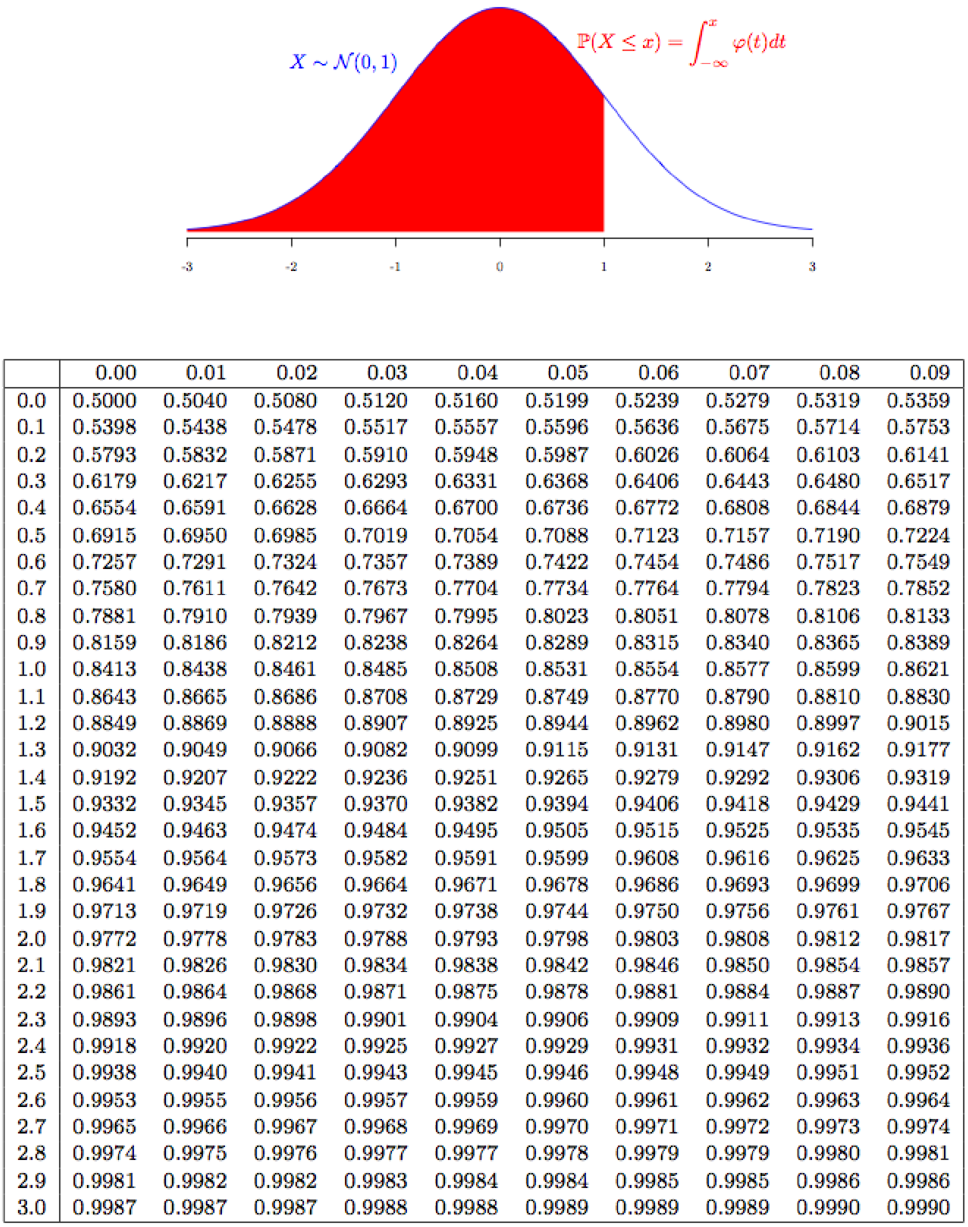
Exercise 1. Small questions

1. Give a possible sample of size 4 from each of the following populations:
    - all daily newspapers published in the world,
    - all grades of students at the probability course,
    - all distances that may result when you throw a football,
    - all possible daily numbers of cars going through a crossing.
2. In 1882, Michelson and Newcomb measured the traveling time of light going from and to their lab through a mirror. Their first measurements were: 28, 26, 33, 24, 34, -44, 27, 16, 40, -2, 29, 24, 21, 25 ($*0.001 + 24.8$ in millionths of a second). Why are these measurements not identical? How do we model this variability in statistics?
3. Are the following statistical models identifiable
    - $\big((\mathcal{X} = \mathbb{R}, \mathcal{B}(\mathbb{R})), \{\mathcal{N}(\mu + a, \sigma^2), (\mu, a, \sigma) \in \mathbb{R} \times \mathbb{R} \times \mathbb{R}_+\}\big)$,
    - $\big((\mathcal{X} = \mathbb{R}, \mathcal{B}(\mathbb{R})), \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}\big)$,
    - $\big((\mathcal{X} = \mathbb{R}, \mathcal{B}(\mathbb{R})), \{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}\}\big)$?

Solution

1. - Le Monde, The New York Times, The Guardian, El País
    - A, A, D, B
    - 10.45m, 3.23m, 7.53m, 15.74m
    - 102, 84, 26, 73
2. The measurements are not identical because of the uncertainty of the experiment and the measurements, small errors, perturbations happen for each light traveling. We model this variability in statistics thanks to the probability theory (theory of the randomness), saying that these observed values $x_1 = 28$, $x_2 = 26$, $x_3 = 33$, $\dots$ are the realization of i.i.d. random variables X_1, X_2, $\dots$.
3. - no, $\mathcal{N}(\mu_1, \sigma^2) = \mathcal{N}(\mu_2 + \delta - \delta, \sigma^2)$ for any real δ. So that the same distribution is obtained with $\mu = m + \delta$, $a = -\delta$, for any real δ.
    - yes, since the first moment and the second moment completely identify the two parameters. Assume that $\mathcal{N}(\mu_1, \sigma_1^2) = \mathcal{N}(\mu_2, \sigma_2^2)$ then $\mu_1 = \mathsf{E}_{\mathcal{N}(\mu_1,\sigma_1^2)}(X) = \mathsf{E}_{\mathcal{N}(\mu_2,\sigma_2^2)}(X) = \mu_2$ and $\sigma_1^2 = Var_{\mathcal{N}(\mu_1,\sigma_1^2)}(X) = Var_{\mathcal{N}(\mu_2,\sigma_2^2)}(X) = \sigma_2^2$.
    - No, $\mathcal{N}(\mu, \sigma^2) = \mathcal{N}(\mu, (-\sigma)^2)$

Exercise 2. Gaussian measurements
For this exercise, you can use the table of the cdf of the standard normal distribution in Figure 1. The measurement of atmospheric ozone concentration (in $\mu g/m^3$) is modeled by a random variable X with distribution $\mathcal{N}(m, \sigma^2)$ with $\sigma^2 = 3.1$.


$X \sim \mathcal{N}(0,1)$
$\mathbb{P}(X \leq x) = \int_{-\infty}^{x} \varphi(t)dt$
-3
-2
-1
0
1
2
3

	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990

Figure 1: Table of the cdf of the standard normal distribution

1. Write the statistical model.
2. In many applications, data are often modeled with the Normal distribution, while often the observed values are by definition positive (e.g. weight, size, speed, duration). Can you explain why?
3. What are the units of m and σ? What do m and σ represent?
4. The ozone concentration is considered dangerous for humans when it is greater than $180\mu g/m^3$.
   - (a) Assuming that $m = 178$, what is the probability that the measurement is greater than 180? Comment the result.
   - (b) Assuming that $m = 183$, what is the probability that the measurement is smaller than 180? Comment the result.

(c) Assuming that $m = 180$, find a real number δ such that the probability $P(180-\delta \le X \le 180+\delta)$ is bigger than 95%.

5. Are the three last questions statistical or probabilistic questions? Find a question that a statistician could ask from this experience.

6. Some day, we make some measurements and we assume that this day the ozone concentration is $178\mu g/m^3$ (yet the experimenter doesn't know this concentration otherwise he wouldn't need measurements).

   (a) Compute the probability that a unique measurement is greater than 180?

   (b) What is the probability that the mean of three measurements is greater than 180 ?

   (c) How many measurements are necessary for the probability that the mean of these measurements is greater than 180 being less than 1%?

Solution

1. $\big((\mathbb{R}, \mathcal{B}(\mathbb{R}), \{\mathcal{N}(m, \sigma^2), \theta = (m, \sigma) \in \mathbb{R} \times \mathbb{R}_+\}\big)$ we assume that the measurement of the atmospheric ozone concentration X is distributed from a Gaussian distribution $\mathcal{N}(m, \sigma^2)$ for some unknown $\theta = (m, \sigma) \in \mathbb{R} \times \mathbb{R}_+$. Here the population, all the possible measurements, is supposed to be $\mathbb{R}$, and we have a sample, one measurement x which is a realization of X.

2. In many applications, the data are concentrated around one value and look symmetric. The normal distribution being unimodal and symmetric is then a good candidate.
   In this exercise the population is supposed to be $\mathbb{R}$, the support of the Normal distribution is also $\mathbb{R}$, while the concentration cannot be negative. Worse, for any m, σ, the probability that X is negative is always positive. Yet given that the Gaussian distribution has light tails this probability is usually not high. For instance, when $m = 180$, for all $\sigma \leqslant 22$,

$$\mathbb{P}(X \leqslant 0) = \mathbb{P}\left(\frac{X-m}{\sigma} \leqslant \frac{-m}{\sigma}\right) = F_{\mathcal{N}(0,1)}(-m/\sigma) \leqslant 10^{-15}.$$

   We have to remember that 'all models are wrong but some are useful'. Then a good model is a model which is a good approximation of the reality and is tractable. You have to find a balance between the complex reality and an easy mathematical model which is useful in practice.

3. m and σ are in $\mu g/m^3$. m represents the real atmospheric concentration and σ the imprecision of the measure.

4. (a) This probability is

$$\begin{aligned} P_{\mathcal{N}(178,3.1)}(X \ge 180) &= P_{\mathcal{N}(178,3.1)}\left(\frac{X-178}{\sqrt{3.1}} \ge \frac{180-178}{\sqrt{3.1}}\right) \\ &= P_{U\sim\mathcal{N}(0,1)}\left(U \ge \frac{2}{\sqrt{3.1}}\right) = 1 - F_{\mathcal{N}(0,1)}\left(\frac{2}{\sqrt{3.1}}\right) \sim 12.74\%, \end{aligned}$$

   where $F_{\mathcal{N}(0,1)}$ is the cdf of a standard normal distribution.

   (b)

$$\begin{aligned} P_{\mathcal{N}(183,3.1)}(X \le 180) &= P_{\mathcal{N}(183,3.1)}\left(\frac{X-183}{\sqrt{3.1}} \le \frac{180-183}{\sqrt{3.1}}\right) \\ &= P_{U\sim\mathcal{N}(0,1)}\left(U \le -\frac{3}{\sqrt{3.1}}\right) = P_{U\sim\mathcal{N}(0,1)}\left(U \ge \frac{3}{\sqrt{3.1}}\right) \\ &= 1 - F_{\mathcal{N}(0,1)}\left(\frac{3}{\sqrt{3.1}}\right) \sim 4.46\%. \end{aligned}$$

(c)

$$\begin{aligned}
P_{\mathcal{N}(180,3.1)}(180-\delta \leq X \leq 180+\delta) &= P_{\mathcal{N}(180,3.1)}\left(\frac{180-\delta-180}{\sqrt{3.1}} \leq \frac{X-180}{\sqrt{3.1}} \leq \frac{180+\delta-180}{\sqrt{3.1}}\right)\\
&= P_{U\sim\mathcal{N}(0,1)}\left(-\frac{\delta}{\sqrt{3.1}} \leq U \leq \frac{\delta}{\sqrt{3.1}}\right)\\
&= 1-2P_{U\sim\mathcal{N}(0,1)}\left(U \geq \frac{\delta}{\sqrt{3.1}}\right) \leq 0.95\\
&\Leftarrow 2F_{\mathcal{N}(0,1)}\left(\frac{3}{\sqrt{3.1}}\right)-1 \sim 4.46\%,
\end{aligned}$$

hence $\delta = 3$ is a suitable choice.

5. The three last questions are probability questions, since we assume that we know the population, $m = ...$ and we search for the properties of a sample from this population. In inferential statistics, we do the opposite. We have a sample and we try to infer from this sample some information about the population. For instance, we could try to infer m, that is the average ozone concentration from a sample x_1 (or $x_1, \ldots, x_n$). One could infer m from the empirical mean of the sample, that $1/n\sum_{i=1}^n x_i$ if he has n observed measurements.

6. (a) The probability that a unique measure is greater than 180 is $1-F_{\mathcal{N}(0,1)}\left(\frac{2}{\sqrt{3.1}}\right) \sim 12.74\%$ (question 2.a).

   (b) Let X_1, X_2 and X_3 be the three measurements. The mean of these three measurements is $Z = \bar{X}_3 = 1/3\sum_{i=1}^3 X_i$, it is distributed according to a normal distribution. Its expectation is $E(Z) = 178$ and its variance $Var(Z) = 3.1/3$. So that Z is distributed from $\mathcal{N}(178, 3.1/3)$. Then the probability that the mean of three measurements is greater than 180 is

$$P_{\mathcal{N}(178,3.1/3)}(Z \geq 180) = P_{\mathcal{N}(178,3.1/3)}\left(\frac{Z-178}{\sqrt{3.1/3}} \geq \frac{2}{\sqrt{3.1/3}}\right) = 1-F_{\mathcal{N}(0,1)}\left(\frac{2\sqrt{3}}{\sqrt{3.1}}\right) \sim 2.44\%.$$

   (c) We assume that we have n measurements $X_1, \ldots, X_n$. The mean of these measurements is the random variable $\bar{X}_n = 1/n\sum_{i=1}^n X_i$. $\bar{X}_n$ is distributed from a normal distribution $\mathcal{N}(178, 3.1/n)$. We search for n such that $P\left(\bar{X}_n \geq 180\right) \leq 0.01$.

$$P_{\mathcal{N}(178,3.1/3)}\left(\bar{X}_n \geq 180\right) = P_{\mathcal{N}(178,3.1/3)}\left(\frac{\bar{X}_n-178}{\sqrt{3.1/n}} \geq \frac{180-178}{\sqrt{3.1/n}}\right) = 1-F_{\mathcal{N}(0,1)}\left(\frac{2\sqrt{n}}{\sqrt{3.1}}\right).$$

   From the table, we search n such that $\frac{2\sqrt{n}}{\sqrt{3.1}} \geq 2.33$. So that $n \geq 5$.

Exercise 3. We are in front of a black urn which contains N balls which are numbered from 1 to N. We don't know N the number of balls but we can draw as many balls as we wish if we put it in the urn before drawing another one.

1. Write the statistical model.

2. How can you guess the value of N with the observed numbers $x_1, \ldots, x_n$ during the sampling?

3. What is the distribution of the greatest number $\hat{N} = \max(X_1, \ldots X_n)$?

4. How many balls do you want to draw? Help: you can compute the probability that $\hat{N} = N$.

Solution

1. If we draw one ball and see the number X_1, the statistical model associated is

$$(\mathbb{N}, \mathcal{P}(\mathbb{N}), P_N, N \in \mathbb{N})$$

where $\mathcal{P}(\mathbb{N})$ is the power set of $\mathbb{N}$ and P_N, for $N \in \mathbb{N}$ is the uniform distribution on $\{1, \ldots, N\}$.
If we draw n balls with numbers $X_1, \ldots, X_n$, the associated statistical model is

$$\big(\mathbb{N}^n, (\mathcal{P}(\mathbb{N}))^n, (P_N)^{\otimes n}, N \in \mathbb{N}\big)$$

where $\mathcal{P}(\mathbb{N})$ is the power set of $\mathbb{N}$ and P_N, for $N \in \mathbb{N}$ is the uniform distribution on $\{1, \ldots, N\}$. In the following we consider that we have observed n numbers $x_1, \ldots, x_n$ which are realizations of $X_1, \ldots, X_n$.

2. $\hat{N}_{x_1,\ldots,x_n} = \max(x_1, \ldots, x_n)$. Here we want to find the distribution of a maximum, we then use the cdf to characterize the distribution of $\hat{N}$. We assume that the observations come from P_N for some $N \in \mathbb{N}$. Let $t \in \mathbb{R}$ (common error: $t \in \mathbb{N}$ while the cdf is defined on $\mathbb{R}$, make a picture if necessary),

$$\begin{aligned} P_N(\hat{N} \leq t) &= P_N(\forall 1 \leq i \leq n,\ X_i \leq t) \\ &= \big(P_N(X_1 \leq t)\big)^n = \begin{cases} \left(\frac{\lfloor t \rfloor}{N}\right)^n & \text{if } 0 \leq t < N \\ 0 & \text{if } t < 0 \\ 1 & \text{otherwise.} \end{cases} \end{aligned}$$

3. We want to draw as many balls as we can because for all N,

$$P_N(\hat{N} = N) = P_N(\hat{N} \leq N) - P(\hat{N} \leq N - 1) = 1 - \left(\frac{N-1}{N}\right)^n \xrightarrow{n \to \infty} 1.$$

2 To train

Exercise 4. Heart attack and aspirin
Devore, in his book *Probability and statistics* writes the following:
‘ An article in the New York times (Jan.27, 1987) reported that heart attack risk could be reduced by taking aspirin. This conclusion was based on a designed experiment involving both a control group of individuals that took a placebo (...) and a treatment group that took aspirin (...). Of the $11,034$ individuals in the control group, 189 subsequently experienced heart attacks, whereas only 104 of the $11,037$ in the aspirin group had heart attack. The incidence rate of heart attacks in the treatment group was only about half that in the control group. One possible explanation for this result is chance variation–that aspirin really doesn’t have the desired effect and the observed difference is just typical variation in the same way that tossing two identical coins would usually produce different numbers of heads. However, in this case, inferential methods suggest that chance variation by itself cannot adequately explain the magnitude of the observed difference.’
Can you formalize statistically this description?

Solution In this experience, there are two samples. One sample comes from the control group, i.e. the hypothetical population of individuals taking placebo. Let $x_i = 1$ if the i-th individual of the control group had a heart attack and $x_i = 0$ otherwise for i between 1 and $11,034 = n_x$. These observed values are supposed to be realizations of i.i.d. random variables $X_1, \ldots, X_{n_x}$, with values in $\mathcal{X} = \{0, 1\}$. So that, X_i are distributed from a Bernoulli distribution with parameter θ_x, which represents the proportion of individuals experiencing heart attack in the hypothetical control population.
The other sample comes from the treatment group. It is a sample of the hypothetical population of individuals taking some dose of aspirin.Let $y_i = 1$ if the i-th individual of the treatment group had a heart attack and $y_i = 0$ otherwise for i between 1 and $11,037 = n_y$. These observed values are supposed to be realizations of i.i.d. random variables $Y_1, \ldots, Y_{n_y}$, with values in $\mathcal{Y} = \{0, 1\}$. So that, Y_i are distributed from a Bernoulli distribution with parameter θ_y, which represents the proportion of individuals experiencing heart attack in the hypothetical treatment population.
One could conclude that heart attack risk could be reduced by taking aspirin if $\theta_x > \theta_y$. But we don’t have access to these parameters describing hypothetical populations. Researchers have to infer information on these two parameters θ_x and θ_y thanks to the samples. On the two samples we know that $189 = \sum_{i=1}^{n_x} x_i$ individuals of the control group experienced a heart attack and $104 = \sum_{i=1}^{n_y} y_i$ individuals of the treatment group experienced a heart attack. Intuitively, we can approximate θ_x with $\bar{x} = \frac{1}{n_x}\sum_{i=1}^{n_x} x_i$ and θ_y with $\bar{y} = \frac{1}{n_y}\sum_{i=1}^{n_y} y_i$.
Saying that ‘one possible explanation of this result is chance variation (...)’ means that maybe $\theta_x = \theta_y$,

and under this assumption, the probability that $\bar{X} = \frac{1}{n_x}\sum_{i=1}^{n_x} X_i \neq \bar{Y} = \frac{1}{n_y}\sum_{i=1}^{n_y} Y_i$ is nonzero (even has high probability). Saying that 'inferential methods suggest that chance variation by itself cannot adequately explain the magnitude of the observed difference' means that using inferential tools, that we will study later (hypotheses tests), we can reject the fact $\theta_x = \theta_y$ because the probability, under $\theta_x = \theta_y$, that the random variables is as or more extreme than the observed one, i.e. $\bar{X}/\bar{Y} \geqslant \bar{x}/\bar{y}$, is too small.

Exercise 5. Survival analysis and identifiability

We study a system which works as long as two machines work. The durations X_1 and X_2 of these two machines follow exponential distributions with parameters λ_1 and λ_2 : $\mathbb{P}(X_i > x) = \exp(-\lambda_i x)$.

X_1 and X_2 are assumed independent.

1. Compute the probability of the following event: "the system does not break down before time t".

   Deduce from this the probability distribution of Z, the duration of the system.

   Compute the probability of "The failure is due to machine 1".

2. Let $I = 1$ if the failure is due to machine 1 and $I = 0$ otherwise. Compute $\mathbb{P}(\{Z > t\} \cap \{I = \delta\})$, for any $t \geq 0$ and $\delta \in \{0, 1\}$. Deduce from this that Z and I are independent.

3. We are now provided with n identical systems (as in bullet '1.') which are independent from each other. Their durations are observed and given by $Z_1, ..., Z_n$.

   (a) Write the corresponding statistical model. Are the parameters λ_1 and λ_2 identifiable?

   (b) Assume now that we observe these duration and the corresponding reasons of failures $I_1, ..., I_n$, where $I_i \in \{0, 1\}$. Write the corresponding statistical model. Are the parameters λ_1 and λ_2 identifiable?

Solution

1. Let Z be the random variable modeling the duration of the system. By independence of X_1 and X_2,

$$\mathbb{P}(Z > t) = \mathbb{P}(\{X_1 > t\} \cap \{X_2 > t\}) = \mathbb{P}(\{X_1 > t\})\mathbb{P}(\{X_2 > t\}) = e^{-\lambda_1 t}e^{-\lambda_2 t} = e^{-(\lambda_1+\lambda_2)t}.$$

   The distribution of the duration of the system is an Exponential distribution with parameter $\lambda_1 + \lambda_2$. The probability that the failure is due to machine one is

$$\mathbb{P}(X_2 > X_1) = \int_0^{+\infty}\int_{x_1}^{+\infty} e^{-\lambda_1 x_1}e^{-\lambda_2 x_2}\lambda_1\lambda_2 \mathrm{d}x_2\mathrm{d}x_1 = \int_0^{+\infty} \lambda_1 e^{-\lambda_1 x_1} \underbrace{\int_{x_1}^{\infty} \lambda_2 e^{-\lambda_2 x_2}\mathrm{d}x_2}_{\left[-e^{-\lambda_2 x_2}\right]_{x_2=x_1}^{x_2=\infty}=e^{-\lambda_2 x_1}} \mathrm{d}x_1$$

$$= \int_0^{\infty} \lambda_1 e^{-(\lambda_1+\lambda_2)x_1}\mathrm{d}x_1 \overset{(*)}{=} \lambda_1 \underbrace{\left[\frac{e^{-(\lambda_1+\lambda_2)x_1}}{-(\lambda_1+\lambda_2)}\right]_{x_1=0}^{x_1=\infty}}_{\frac{1}{\lambda_1+\lambda_2}} = \frac{\lambda_1}{\lambda_1+\lambda_2}.$$

2. Using that $\{Z > t\} \cap \{I = 1\} = \{X_1 > t\} \cap \{X_2 > t\} \cap \{I = 1\} = \{X_1 > t\} \cap \{X_2 > X_1\}$, we get

$$\mathbb{P}(\{Z > t\} \cap \{I = 1\}) = \int_t^{+\infty}\int_{x_1}^{+\infty} e^{-\lambda_1 x_1}e^{-\lambda_2 x_2}\lambda_1\lambda_2\mathrm{d}x_2\mathrm{d}x_1 \overset{(*)}{=} \lambda_1 \underbrace{\left[\frac{e^{-(\lambda_1+\lambda_2)x_1}}{-(\lambda_1+\lambda_2)}\right]_{x_1=t}^{x_1=\infty}}_{\frac{e^{-(\lambda_1+\lambda_2)t}}{\lambda_1+\lambda_2}}$$

$$= \frac{\lambda_1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)t} = \mathbb{P}(I = 1)\mathbb{P}(Z > t).$$

   Hence,

$$\mathbb{P}(\{Z > t\} \cap \{I = 0\}) = \mathbb{P}(Z > t) - \mathbb{P}(\{Z > t\} \cap \{I = 1\}) = e^{-(\lambda_1+\lambda_2)t} - \frac{\lambda_1}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)t}$$

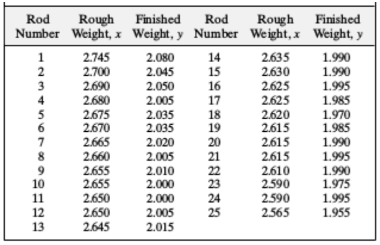
$$= \frac{\lambda_2}{\lambda_1+\lambda_2}e^{-(\lambda_1+\lambda_2)t} = \mathbb{P}(I = 0)\mathbb{P}(Z > t).$$

   Then Z and I are independent random variables.

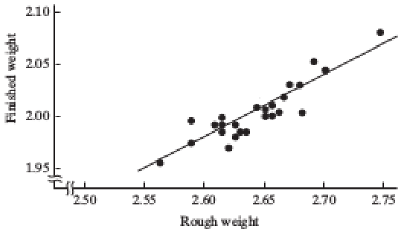
3. (a) The statistical model is defined by the probability space $(0, +\infty)^n$ (endowed with the Borel σ-field) and the family of distributions $(\mathcal{E}(\lambda_1 + \lambda_2))^{\otimes n}$, so that Z_i are i.i.d. Exponential with parameter $\lambda_1 + \lambda_2$, parametrized with the parameters $\lambda_1 \in \mathbb{R}_+$ and $\lambda_2 \in \mathbb{R}_+$ ($\theta = (\lambda_1, \lambda_2) \in \mathbb{R}_+^2$).
The parameters λ_1 and λ_2 are not identifiable because they are only known through their sum. For instance, $\theta = (\lambda_1, \lambda_2)$ and $\theta' = (1/2\lambda_1, \lambda_2 + 1/2\lambda_1)$ are unequal but leads to the same distribution.

   (b) The statistical model is defined by the probability space $((0, +\infty) \times \{0, 1\})^n$ and the family of distributions $(\mathcal{E}(\lambda_1 + \lambda_2) \otimes \mathcal{B}(\lambda_1/(\lambda_1 + \lambda_2))^{\otimes n}$, parametrized with the parameters $\lambda_1 \in \mathbb{R}_+$ and $\lambda_2 \in \mathbb{R}_+$ ($\theta = (\lambda_1, \lambda_2) \in \mathbb{R}_+^2$).
The parameters λ_1 and λ_2 are now identifiable. For instance by computing the expectation of Z_1, we recover $\lambda_1 + \lambda_2$ and by computing the expectation of I_1 we obtain $\lambda_1/(\lambda_1 + \lambda_2)$. Then λ_1 and λ_2 can be deduced.

Exercise 6. Linear regression (Larsen and Marx *An introduction to mathematical statistics and its applications*)
A manufacturer of air conditioning units is having assembly problems due to the failure of a connecting rod to meet finished-weight specifications. Too many rods are being completely tooled, then rejected as overweight. To reduce that cost, the company's quality-control department wants to quantify the relationship between the weight of the finished rod, and that of the rough casting. Castings likely to produce rods that are too heavy can then be discarded before undergoing the final (and costly) tooling process. The following Figures represent the dataset that the company have.



Rod Number	Rough Weight, x	Finished Weight, y	Rod Number	Rough Weight, x	Finished Weight, y
1	2.745	2.080	14	2.635	1.990
2	2.700	2.045	15	2.630	1.990
3	2.690	2.050	16	2.625	1.995
4	2.680	2.005	17	2.625	1.985
5	2.675	2.035	18	2.620	1.970
6	2.670	2.035	19	2.615	1.985
7	2.665	2.020	20	2.615	1.990
8	2.660	2.005	21	2.615	1.995
9	2.655	2.010	22	2.610	1.990
10	2.655	2.000	23	2.590	1.975
11	2.650	2.000	24	2.590	1.995
12	2.650	2.005	25	2.565	1.955
13	2.645	2.015			



1. Propose a statistical model.
2. Propose a statistical question that the company want to answer?

Solution

1. The company has a sample of $n = 25$ rods. It is a sample of a hypothetical population of all possible rods/rough casting. For each rod $i \leqslant n$, they have measured x_i the rough casting weight and y_i the final rod weight. We assume that $(x_1, y_1), \ldots, (x_n, y_n)$ are realizations of i.i.d. bivariate random variables (X_i, Y_i) for $i \leqslant n$. Looking at the scatter plot of the sample, we observe that the points (x_i, y_i) seem to be aligned. We may think that the final weight is a linear function of the rough casting weight. So that $Y_i = aX_i + b$, this model would lead to points perfectly aligned; which is not the case because of noise. So we add some noise to this model, and we model the relationship between the two weights as follows: $Y_i = aX_i + b + \epsilon_i$ with ϵ_i a random variable distributed from some centered distribution P_θ, for instance the normal distribution $\mathcal{N}(0, \sigma^2)$. X_i is distributed from some distribution Q that we are not interested in and given X_i, Y_i is distributed from $\mathcal{N}(aX_i + b, \sigma^2)$. The parameters describing the model are then $\theta = (a, b, \sigma^2) \in \mathbb{R}^2 \times \mathbb{R}_+$ and Q.

2. The company wants 'to quantify the relationship between' the two weights, that is to say, it wants to know the parameters describing the model, i.e. to estimate a, b and σ.

3 To go further

Exercise 7. Autoregressive model
Let us consider the following observation $Z = (X_1, ..., X_n)$, where X_i are generated through an autoregressive model:

$$X_i = \theta X_{i-1} + \xi_i, \quad i = 1, ..., n, \quad X_0 = 0,$$

where ξ_i are i.i.d. $\mathcal{N}(0, \sigma^2)$ and $\theta \in \mathbb{R}$. Write the statistical model given by the observation Z.

Solution

$$X_0 = 0, \quad X_i = \theta X_{i-1} + \xi_i, \quad \xi_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$$

Statistical model

- Probability space: $\mathbb{R}^n$ (endowed with the Borel σ-field).
- The family of distribution $P_{\theta,\sigma}$ is parametrized by $\theta \in \mathbb{R}$ and $\sigma^2 \in \mathbb{R}_+$. Under $P_{\theta,\sigma}$,

$$Z^T = \begin{pmatrix} X_1 \\ X_2 \\ \\ \vdots \\ \\ X_n \end{pmatrix} = \begin{pmatrix} 1 & 0 & & & 0 & 0 \\ \theta & 1 & & & & 0 \\ \theta^2 & \theta & 1 & \ddots & & \\ & \ddots & \ddots & \ddots & & \\ \theta^{n-2} & & & \theta & 1 & 0 \\ \theta^{n-1} & \theta^{n-2} & & \theta^2 & \theta & 1 \end{pmatrix} \begin{pmatrix} \xi_1 \\ \xi_2 \\ \\ \vdots \\ \\ \xi_n \end{pmatrix} =: A_\theta \xi,$$

where ξ is a Gaussian vector with distribution $\mathcal{N}(0, \sigma^2 I_n)$. So that Z is a Gaussian vector with distribution $\mathcal{N}(0, \sigma^2 A_\theta A_\theta^T)$ which admits the following density (w.r.t. the Lebesgue measure)

$$z \in \mathbb{R}^n \mapsto \frac{1}{\sqrt{2\pi\sigma^n}} \frac{1}{\sqrt{\det(\Gamma_\theta)}} \exp\left(-\frac{1}{2\sigma^2} z^T \Gamma_\theta^{-1} z\right),$$

where

$$\Gamma_\theta = (A_\theta A_\theta^T)^{-1} = \begin{pmatrix} 1+\theta^2 & -\theta & 0 & & 0 & 0 \\ -\theta & 1+\theta^2 & -\theta & 0 & & 0 \\ 0 & -\theta & 1+\theta^2 & -\theta & & \\ & \ddots & \ddots & \ddots & \ddots & \\ 0 & & & -\theta & 1+\theta^2 & -\theta \\ 0 & 0 & & 0 & -\theta & 1+\theta^2 \end{pmatrix}.$$

Exercise 8. Bayesian modeling

An exam has 10 questions, each with 3 possible answers. Assume that students, who are prepared correctly, answer correctly each question with probability 0.8 and that the other students answer at random. The score S of a student is the sum of the points when 1 is attributed for each correct answer (0 otherwise). We would like to know if the student is prepared.

1. Characterize the distribution of S if the student is prepared and if he is not.
2. Give the statistical model associated to this experiment.

The previous years, 70% of the students were prepared.

3. Which prior distribution would you choose? (this defines θ, see 5.)
4. What is the posterior probability that the student is prepared given that its score is 5? Deduce the posterior distribution given the observation $S = 5$?

Another exam has 15 questions.

5. What is the posterior distribution of θ given a score 8?

We now consider a exam with n questions.

6. What is the posterior distribution of θ given a score s?
7. What happens when s increases?

8. Consider a score proportional to n: $s = cn$ with $c \in [0, 1]$. Compute the posterior distribution given a score $s = cn$. What happens when n tends to $+\infty$?

Solution

1. If the student is prepared, S is distributed from a Binomial distribution $Bin(10, 0.8)$, since it is a sum of independent Bernoulli $\mathcal{B}(0.8)$ random variables. If the student is not prepared, S is distributed from a Binomial distribution $Bin(10, 1/3)$.

2. $(\{0, \dots, 10\}, \mathcal{P}(\{0, \dots, 10\}), B_\theta, \theta \in \{0, 1\})$ where B_0 is a Binomial distribution with parameter $(10, 0.8)$ and B_1 is a Binomial distribution with parameter $(10, 1/3)$.

3. We choose a Bernoulli $\mathcal{B}(0.7)$ distribution as prior distribution Π for $\theta \in \{0, 1\}$.

4. We assume that we observe $S = 5$. We search for the posterior probability that $\theta = 1$ given that we observe $S = 5$. By the Bayes' theorem

$$\begin{aligned}\Pi(\theta = 1|S = 5) &= \frac{P(\theta = 1 \text{ and } S = 5)}{P(S = 5)} = \frac{P(\theta = 1 \text{ and } S = 5)}{P(\{S = 5 \text{ and } \theta = 1\} \cup \{S = 5 \text{ and } \theta = 0\})} \\ &= \frac{P(\theta = 1 \text{ and } S = 5)}{P(\{S = 5 \text{ and } \theta = 1\}) + P(\{S = 5 \text{ and } \theta = 0\})} \\ &= \frac{P(S = 5|\theta = 1)P(\theta = 1)}{P(S = 5|\theta = 1)P(\theta = 1) + P(S = 5|\theta = 0)P(S = 0)},\end{aligned}$$

where P is the joint distribution of the parameter θ and S. We know that the distribution of S given $\theta = 1$ is a Binomial $Bin(10, 0.8)$ and the distribution of S given $\theta = 0$ is a Binomial $Bin(10, 1/3)$ and θ is distributed from a Bernoulli $\mathcal{B}(0.7)$. So

$$\begin{aligned}\Pi(\theta = 1|S = 5) &= \frac{P(S = 5|\theta = 1)P(\theta = 1)}{P(S = 5|\theta = 1)P(\theta = 1) + P(S = 5|\theta = 0)P(\theta = 0)} = \frac{B_1(5)\Pi(1)}{B_1(5)\Pi(1) + B_0(5)\Pi(0)} \\ &= \frac{\binom{10}{5}0.8^5 0.2^5 0.7}{\binom{10}{5}0.8^5 0.2^5 0.7 + \binom{10}{5}(1/3)^5(2/3)^5 0.3} = 0.311\end{aligned}$$

and the posterior distribution given a score of 5 is Bernoulli distribution with parameter $\Pi(\theta = 1|S = 5) = 0.311$.

5. The new statistical model is $(\{0, \dots, 15\}, \mathcal{P}(\{0, \dots, 15\}), B_\theta, \theta \in \{0, 1\})$ where B_0 is a Binomial distribution with parameter $(15, 0.8)$ and B_1 is a Binomial distribution with parameter $(15, 1/3)$. We keep the same prior distribution. The posterior probability that the student is prepared given that $S = 8$ is

$$\begin{aligned}\Pi(\theta = 1|S = 8) &= \frac{P(S = 8|\theta = 1)P(\theta = 1)}{P(S = 8|\theta = 1)P(\theta = 1) + P(S = 8|\theta = 0)P(\theta = 0)} \\ &= \frac{\binom{15}{8}0.8^8 0.2^7 0.7}{\binom{15}{8}0.8^8 0.2^7 0.7 + \binom{15}{8}(1/3)^8(2/3)^7 0.3} = 0.3597.\end{aligned}$$

Then the posterior distribution given a score of 5 is Bernoulli distribution with parameter $\Pi(\theta = 1|S = 8) = 0.3597$.

6. The new statistical model is $(\{0, \dots, n\}, \mathcal{P}(\{0, \dots, n\}), B_\theta, \theta \in \{0, 1\})$ where B_0 is a Binomial distribution with parameter $(n, 0.8)$ and B_1 is a Binomial distribution with parameter $(n, 1/3)$. We keep the same prior distribution. The posterior probability that the student is prepared given that $S = s$ is

$$\begin{aligned}\Pi(\theta = 1|S = s) &= \frac{P(S = s|\theta = 1)P(\theta = 1)}{P(S = s|\theta = 1)P(\theta = 1) + P(S = s|\theta = 0)P(\theta = 0)} \\ &= \frac{\binom{n}{s}0.8^s 0.2^{n-s} 0.7}{\binom{n}{s}0.8^s 0.2^{n-s} 0.7 + \binom{n}{s}(1/3)^s(2/3)^{n-s} 0.3} = \frac{1}{1 + (1/8)^s(10/3)^n 0.3/0.7}.\end{aligned}$$

Then the posterior distribution given a score of 5 is Bernoulli distribution with parameter $\Pi(\theta = 1|S = s) = \frac{1}{1+(1/8)^s(10/3)^n 0.3/0.7}$.

7. We use the previous computation. Then, the posterior probability that the student is prepared given that its score is s increases, when s increases.

8. When $s = cn$,

$$\Pi(\theta = 1|S = cn) = \frac{1}{1 + (10/(3.8^c)^n 0.3/0.7} \xrightarrow{n\to\infty} \begin{cases} 0 & \text{if } c < \frac{\log 10/3}{\log 8} \\ 1 & \text{if } c > \frac{\log 10/3}{\log 8} \\ 0.7 & \text{if } c = \frac{\log 10/3}{\log 8} \end{cases}$$

with $\frac{\log 10/3}{\log 8} \sim 0.58$, using that $1R\frac{10}{3\times 8^c} \Leftrightarrow cR\frac{\log\left(\frac{10}{3}\right)}{\log(8)}$ where $R \in \{>, <, =\}$.