Statistiques Appliquées

Luc Deneire

1 TD 1 : Descriptive statistics

1.1 Plot data

- 8 measurements of a bicycle wheel are given, in mm : 700.01, 700.03, 700.15, 700.00, 700.05, 700.02, 700.05 and 700.04. Compute the sample mean and the sample standard deviation. Plot a dot diagram and comment on the data.
- Temperature measurements are as following : 84, 49, 61, 40, 83, 67, 45, 66, 70, 69, 80, 58, 68, 60, 67, 72, 73, 70, 57, 63, 70, 78, 52, 67, 53, 67, 75, 61, 70, 81, 76, 79, 75, 76, 58, 31.

  Compute the sample mean and sample standard deviation. Set aside the most extreme value (31) and recompute. Comment on the difference.

Solution 1.1

- Mean : 700.04375, Standard deviation (is $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2}$) = 0.04658

  A dot plot could be :

  ```
  ------------------mean
  .---.---.---.---.---.--------------------.
  .----------------------.
  ```

  There is probably an "outlier"

a) $\bar{x}$ = 65.85
   s = 12.16

b) Dot Diagram

```
                                          :  :
  .        .    .   .  ..   .: .: .  .:..: .. :: .... ..
-+---------+---------+---------+---------+---------+-----temp
30        40        50        60        70        80
```

c) Removing the smallest observation (31), the sample mean and standard deviation become
   $\bar{x}$ = 66.86
   s = 10.74

1.2 Plots for large data sets

- Given the following data:

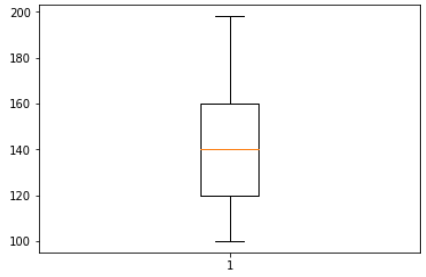
```
[140 100 116 100 126 160 146 121 157 160 185 159 103 162 130 123 198 130
159 113 167 111 120 140 159 128 181 108 123 181 149 101 147 163 106 183
162 155 120 136 129 184 169 196 120 153 101 130 122]
```

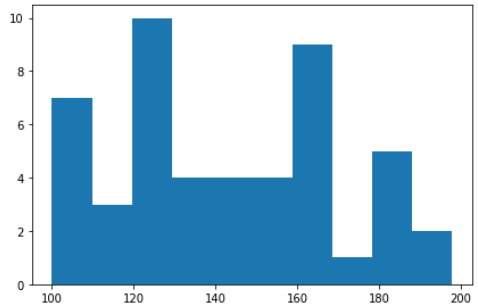
Trace a “stem and leaves” representations, a histogram and the box plot of these data.

Solution 1.2

```
[140 100 116 100 126 160 146 121 157 160 185 159 103 162 130 123 198 130
 159 113 167 111 120 140 159 128 181 108 123 181 149 101 147 163 106 183
 162 155 120 136 129 184 169 196 120 153 101 130 122]
mean of data is :  141.46938775510205
first quartile :  120.0  median :  140.0 thirds quartile :  160.0
```

Out[57]:
```
(array([ 7.,  3., 10.,  4.,  4.,  4.,  9.,  1.,  5.,  2.]),
 array([100. , 109.8, 119.6, 129.4, 139.2, 149. , 158.8, 168.6, 178.4,
        188.2, 198. ]),
 <a list of 10 Patch objects>)
```





- Given the following data:

```
[14.67 -0.91  7.21 -5.73  5.21 -1.4  12.47  4.28  2.08  8.22 -5.81 -0.96
-2.35  8.88  1.22  3.73  7.49  7.73  1.54  4.24  2.82  5.77  1.25 -7.76
20.63  6.94 -3.68  5.56 -4.87 -2.61  3.49  4.37  4.7   3.07 -2.77  1.91
 2.72  0.79  7.34  6.08 -3.62 11.34  0.73 -1.02 11.37  7.37  0.86  5.87
 1.16  4.91  1.55  3.1  12.26  5.31  4.01 -0.69 11.01  2.78 13.38  2.14
 5.1  13.74  7.03  3.94 -0.45  4.21 -2.87 -2.67  2.04  7.38  4.07  5.58
 9.02  3.44 -2.92  3.23  0.1  -5.42  1.49 15.45  2.56 -4.59 10.45  0.15
 5.82  9.6   1.19  4.58 -7.54 -8.88  8.35  5.06 11.84  0.38  2.25 -2.42
-0.32  7.82  2.96  8.57]
```

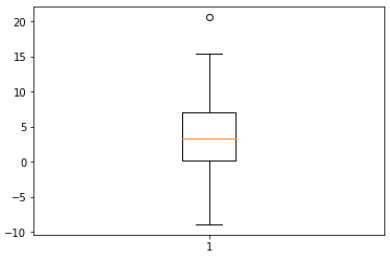
Trace a “stem and leaves” representation, a histogram and the box plot of these data.

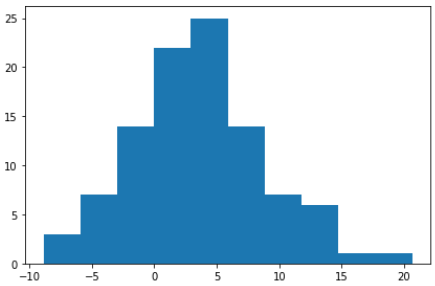
Could these data be the sample of a normal r.v., which mean and standard deviation would it have ?

Solution 1.3

```
[14.67 -0.91  7.21 -5.73  5.21 -1.4  12.47  4.28  2.08  8.22 -5.81 -0.96
 -2.35  8.88  1.22  3.73  7.49  7.73  1.54  4.24  2.82  5.77  1.25 -7.76
 20.63  6.94 -3.68  5.56 -4.87 -2.61  3.49  4.37  4.7   3.07 -2.77  1.91
  2.72  0.79  7.34  6.08 -3.62 11.34  0.73 -1.02 11.37  7.37  0.86  5.87
  1.16  4.91  1.55  3.1  12.26  5.31  4.01 -0.69 11.01  2.78 13.38  2.14
  5.1  13.74  7.03  3.94 -0.45  4.21 -2.87 -2.67  2.04  7.38  4.07  5.58
  9.02  3.44 -2.92  3.23  0.1  -5.42  1.49 15.45  2.56 -4.59 10.45  0.15
  5.82  9.6   1.19  4.58 -7.54 -8.88  8.35  5.06 11.84  0.38  2.25 -2.42
 -0.32  7.82  2.96  8.57]
mean of data is :  3.5072957567845857
first quartile :  0.14096077811268026  median :  3.3349156466991934 thirds quartile :  7.07

(array([ 3.,  7., 14., 22., 25., 14.,  7.,  6.,  1.,  1.]),
 array([-8.88272168, -5.93140801, -2.98009434, -0.02878067,  2.922533  ,
         5.87384667,  8.82516034, 11.77647401, 14.72778768, 17.67910135,
        20.63041502]),
 <a list of 10 Patch objects>)
```





1.3 Normal probability plot

Given the following data set (from the previous exercice) :

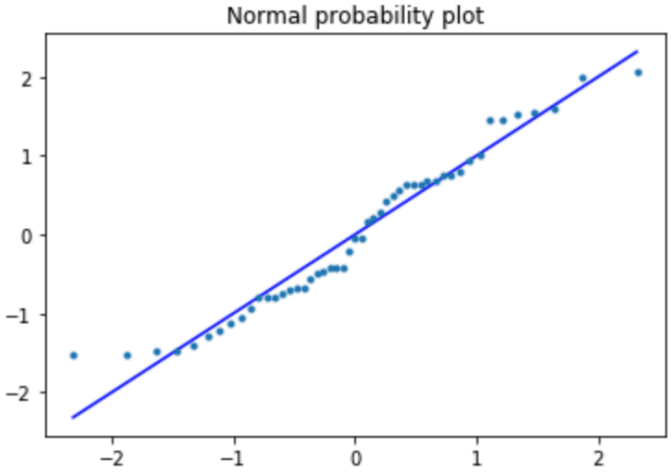
```
[140 100 116 100 126 160 146 121 157 160 185 159 103
162 130 123 198 130 159 113 167 111 120 140 159 128
181 108 123 181 149 101 147 163 106 183 162 155 120
136 129 184 169 196 120 153 101 130 122]
```

From this set, draw a probability plot. Repeat the exercice with the hypothesis that the density was uniform. What can you deduce from this ?

Solution 1.4

A full fledged solution would be :

```
data=np.array([140,100,116,100,126,160,146,121
x=(np.sort(data)-np.mean(data))/np.std(data)
n=data.size;j=np.arange(n)+1;val=((j-0.5)/n)
quantiles = stats.norm.ppf(val,0,1)
fig,ax=plt.subplots()
ax.plot(quantiles,quantiles,'b-')
plt.title('Normal probability plot')
ax.plot(quantiles,x,'.')
plt.show()
```

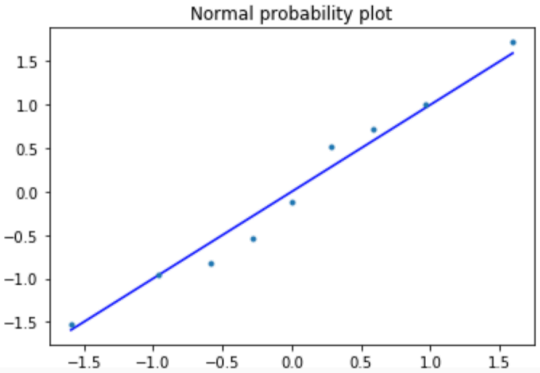


Solution 1.5

But by hand, we have to take a subset of the data :

```
n=x_new.size;j=np.arange(n)+1;val=((j-0.5)/n)
quantiles = stats.norm.ppf(val,0,1)
fig,ax=plt.subplots()
ax.plot(quantiles,quantiles,'b-')
plt.title('Normal probability plot')
ax.plot(quantiles,(x_new-np.mean(x_new))/np.std(x_new),'.')
plt.show()
```

```
[106 120 123 130 140 155 160 167 184]
```



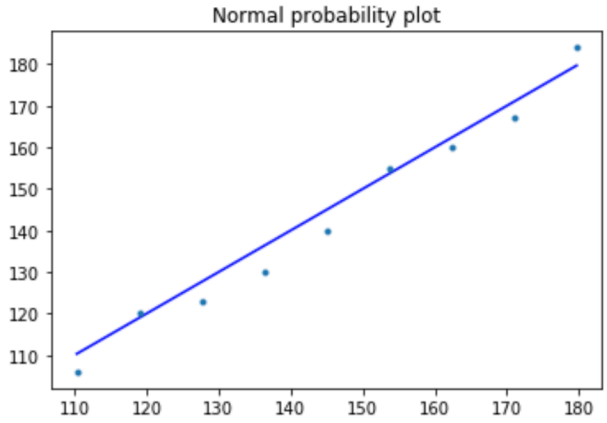
Solution 1.6

If we do it with a uniform hypothesis, we get :

```
quantiles = stats.uniform.ppf(val,106,184-106)
fig,ax=plt.subplots()
print(quantiles)
ax.plot(quantiles,quantiles,'b-')
plt.title('Normal probability plot')
ax.plot(quantiles,(x_new),'.')
plt.show()


```

```
[110.33333333 119.         127.66666667 136.33333333 145.
 153.66666667 162.33333333 171.         179.66666667]
```



and there is no clear cut difference (even with the full set of data)

1.4 Normal probability plot

Given the following data set:

```
[1.35 0.59 0.09 0.17 0.56 1.67 0.14 1.12 0.01 1.84 1.59
 1.73 0.49 0.48 0.07 0.51 0.9  0.57 3.03 0.35]
```

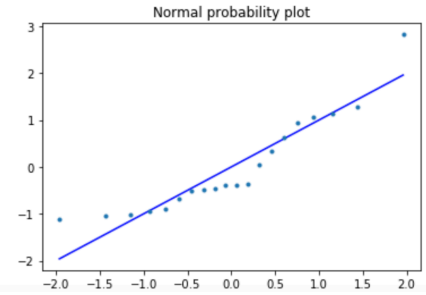
From this set, make a probability plot.

We indicate that this sample is maybe taken from a “Gamma” distribution, whose graphic is given below. Trace the Gamma plot. From this plot, what can you conclude ?

Can you interpret this normal probability plot from the histogram of these data ?

Solution 1.7

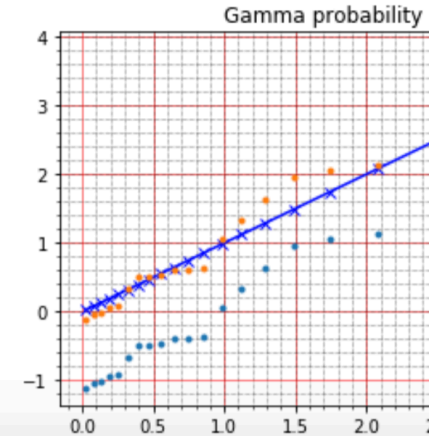
```
data=np.array([1.35 ,0.59 ,0.09, 0.17, 0.56 ,1.67, 0.14, 1.12, 0.01, 1.84, 1.59
 ,1.73, 0.49, 0.48, 0.07, 0.51, 0.9,  0.57, 3.03 ,0.35])
x=(np.sort(data)-np.mean(data))/np.std(data)
n=data.size;j=np.arange(n)+1;val=((j-0.5)/n)
quantiles = stats.norm.ppf(val,0,1)
fig,ax=plt.subplots()
ax.plot(quantiles,quantiles,'b-')
plt.title('Normal probability plot')
ax.plot(quantiles,(x),'.')
plt.show()
```



It looks that for small x_i the values are smaller then the expected values

Solution 1.8

```
x=(np.sort(data)-np.mean(data))/np.std(data)
n=data.size;j=np.arange(n)+1;val=((j-0.5)/n)
quantiles = stats.gamma.ppf(val,1)
fig,ax=plt.subplots()
ax.plot(quantiles,quantiles,'b-x')
plt.title('Gamma probability plot')
ax.plot(quantiles,(x),'.')
ax.plot(quantiles,(x+1),'.')
ax.minorticks_on(); ax.grid(which='major', linestyl
ax.grid(which='minor', linestyle=':', linewidth='0.
plt.show()
```



When plotting for the Gamma Distribution, the smaller values fit much better as well as the larger value.
Note that you have to plot $x + 1$

Given the following data sets :

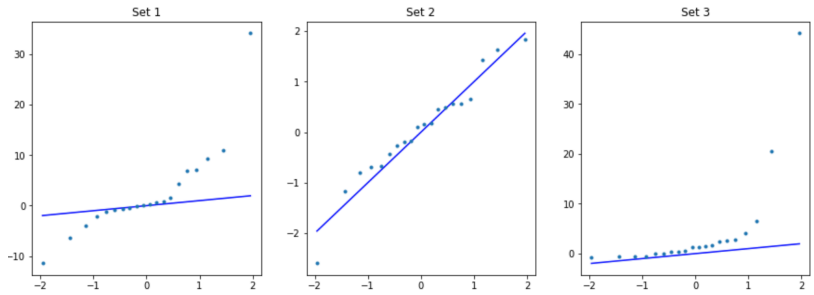
```
Set1 = [-11.45  -6.31  -4.05  -2.22  -1.23  -0.76  -0.68  -0.38  -0.05   0.04
   0.2    0.67   0.84   1.6    4.34   6.88   7.13   9.26  11.02  34.21]

Set2=[-2.6  -1.17 -0.8  -0.69 -0.67 -0.43 -0.27 -0.19 -0.18  0.11  0.16  0.17
  0.46  0.49  0.56  0.57  0.66  1.44  1.64  1.84]

Set3 = [-0.84 -0.56 -0.52 -0.5  -0.1   0.    0.32  0.45  0.56  1.35  1.38  1.42
  1.59  2.46  2.64  2.78  4.01  6.5  20.48 44.19]
```

From their normal probability plot, indicate which set is sampled from a normal probability population. What is the shape of the other plots (in particular, are the laws from which they are taken skewed, do they have a heavy tail ?)

Solution 1.9



The first set has much larger "$x_i s$" compared to the expected values, so it is taken from a heavy tailed distribution.
The second set looks normal.
The third set has larger values for higher quantiles, so it is skewed «to the right».

2 TD2

3 TD 3 : Sampling and Point estimation Estimation

3.1 Life Span

The mean lifespan of an engine is 3000 hours, with a standard deviation of 40 hours, this lifespan has an almost normal distribution. The manufacturer improves his factory, and the lifespan after improvement has a mean of 3035 hours and a standard deviation of 30 hours. We take a sample of n_1 = 16 first generation engines and a sample of n_2 = 25 second generation engines. Which is the probability that the difference in the sample $\bar{X}_2 - \ barX_1$ is more than 25?

Solution 3.1

$$\mathbb{P}\big(\bar{X}_2 - \bar{X}_1 > 25\big) = \mathbb{P}\left(Z > \frac{25 - (\mu_2 - \mu_1)}{\sqrt{\frac{\sigma_1^2}{n_1}\frac{\sigma_2^2}{n_2}}}\right) = \mathbb{P}(Z > -0.857) = 0.8051$$

3.2 Shot accuracy

Two shooters hit their target perfectly on average ($\mathrm{E}[X] = 0$) and with different precision. This precision is given by σ_1 and σ_2. X is assumed to be normal.

We measure 30 shots for shooter 1 and 20 shots for shooter 2. What is the probability that shooter 1 is better than shooter 2, if we measure $s_1^2 = 4mm^2$ and $s_2^2 = 5mm^2$?

Solution 3.2

Shooter 1 is better than shooter 2 if σ_1^2 is smaller than σ_2^2. In the limit, σ_1^2 = σ_2^2, and our statistic is $f_\alpha(n_1 - 1, n_2 - 1) = \frac{s_1^2/\sigma_1^2}{s_2^2/\sigma_2^2} = \frac{s_1^2}{s_2^2} = 4/5$,
In this case, we are interested in $\mathbb{P}(F > 0.8) = \alpha$.
When we look in the tables, we have only ratios larger than 1. Hence we will have to use $f_{(1-\alpha}(n_2 - 1, n_1 - 1) = 1/f_\alpha(n_1 - 1, n_2 - 1) = 1.25$.
Looking at the tables, for $\nu_1 = 20$ (it should be 19, but we don't have it) and $\nu_2 = 29$, we see that $1 - \alpha$ is larger then 25 % and smaller then 50 % %.
Hence, $50\% < \alpha < 75\%$. The actual computer answer gives 71 %.

3.3 Compare estimators

- Given a random sample of size $2n$, taken from a population X of mean μ and variance σ^2.

  We define the following estimators

  $$\hat{X_1} = \frac{1}{2n}\sum_{i=1}^{2n} X_i, \qquad \text{and } \hat{X_2} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

  What is the best estimator, why ?

  Solution 3.3

First, it's easy to prove that both are unbiased estimators.
$\hat{X}_1$ should be the best because :

1. (intuitively), it uses more data
2. It's easy to show that (but students should prove it !) that $\sigma^2_{\hat{X}_1} = \frac{\sigma^2}{2n} < \sigma^2_{\hat{X}_2} = \frac{\sigma^2}{n}$. So $\hat{X}_1$ has a lower variance (and MSE) and is better than $\hat{X}_2$

- Let $\hat{\mathbf{\Theta}_1}$ and $\hat{\mathbf{\Theta}_2}$, be unbiased estimators of θ, with $\sigma^2_{\hat{\mathbf{\Theta}_1}} = 10$ et $\sigma^2_{\hat{\mathbf{\Theta}_2}} = 4$. Which is the best estimator, what is the relative efficiency of $\hat{\Theta}_2$ w.r.t. $\hat{\Theta}_2$?

Solution 3.4

The MSE of $\hat{\Theta}_1$ is 10, the MSE of $\hat{\Theta}_2$ is 4, hence, the second estimator is the best and has a relative efficiency of 10/4 = 2.5 (or 250 percent).

- Let $\hat{\mathbf{\Theta}_1}$, $\hat{\mathbf{\Theta}_2}$ et $\hat{\mathbf{\Theta}_3}$ be three estimators of θ with $\mathrm{E}\left[\hat{\mathbf{\Theta}_1}\right] = \mathrm{E}\left[\hat{\mathbf{\Theta}_2}\right] = \theta, \mathrm{E}\left[\hat{\mathbf{\Theta}_3}\right] \neq \theta, \$\sigma^2_{\hat{\mathbf{\Theta}_1}} = 12, \sigma^2_{\hat{\mathbf{\Theta}_2}} = 10$ et $\mathrm{E}\left[(\hat{\mathbf{\Theta}_3} - \theta)^2\right] = 6$. Compare these estimators, which would you prefer ?

Solution 3.5

Both $\hat{\mathbf{\Theta}_1}$ and $\hat{\mathbf{\Theta}_2}$ are unbiased, but the second has a lower MSE, so we would prefer the second.
$\hat{\mathbf{\Theta}_3}$ has the lowest MSE, but is biased. So if the bias is small enough, we would choose this one. If the bias is too large (and can itself not be estimated), we would choose $\hat{\mathbf{\Theta}_2}$

3.4 Estimation Bias

Let X be a r.v. and $X_1, X_2, \cdots X_n$ a random sample of X. Show that $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ are unbiased estimators of μ and of σ^2.

Solution 3.6

For the mean, it is straightfoward.
For the variance, one possible proof : (others are possible ... on the blackboard I sometimes take other pathes).

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}\mathrm{E}\big[(X_i-\bar{X})^2\big] \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathrm{E}\big[X_i^2\big]-2\mathrm{E}\big[X_i\bar{X}\big]+\mathrm{E}\big[\bar{X}^2\big] \\ &= \frac{1}{n-1}\times n\times[\mathrm{E}\big[X_i^2\big]-2(\mu^2+\frac{\sigma^2}{n})+\mathrm{E}\big[\bar{X}^2\big]] \end{aligned}$$

and

$$\begin{aligned} \mathrm{E}\big[\bar{X}^2\big] &= \mathrm{var}\big[\bar{X}^2\big]+\mathrm{E}\big[\bar{X}\big]^2 \\ &= \frac{1}{n^2}\mathrm{var}\Big[\sum X_i^2\Big]+\mu^2 \\ &= \frac{1}{n^2}\times n\sigma^2+\mu^2 \\ &= \frac{\sigma^2}{n}+\mu^2 \end{aligned}$$

Hence

$$S^2=\frac{n}{n-1}\left(\sigma^2+\mu^2-2(\mu^2+\frac{\sigma^2}{n})+(\frac{\sigma^2}{n}+\mu^2)\right)=\sigma^2$$

3.5 Bow shooting

Two bow shooters are ... shooting a target. We are interested in the arrival point on the horizontal axis, that we will call X_1 and X_2 (for shooters 1 and 2). The two shooters shoot in the middle on average ($\mu_1=\mu_2=0$), but with a different accuracy $\sigma_1^2=a\sigma_2^2$, where $a>1$ is a constant.

we take n_1 and n_2 independent observations of each shooter's result.

- Show that $\hat{\mu}=\alpha\hat{X}_1+(1-\alpha)\hat{X}_2$, with $0<\alpha<1$, is an unbiased estimator of μ.
- Determine the variance of $\hat{mu}$
- DÃ©terminez la valeur de α qui minimise cette variance.
- Supposez $a=4, n_1=2n_2$, quelle valeur de α choisiriez vous ? Quelle serait l'augmentation de l'Ã©cart-type de l'estimateur si on choisissait $\alpha=0.5$?

Solution 3.7

- Straightforward:

$$\mathrm{E}[\hat{\mu}] = \mathrm{E}\Big[\alpha\hat{X}_1 + (1-\alpha)\hat{X}_2\Big] = \alpha\mu + (1-\alpha)\mu = \mu$$

- Independent observations, hence (be sure students can detail the expression) :

$$\sigma_\mu^2 = \alpha^2 \frac{1}{n_1}\sigma_1^2 + (1-\alpha)^2 \frac{1}{n_2}\sigma_2^2 = \sigma_2^2 \frac{\alpha^2.a.n_2 + (1-\alpha)^2 n_1}{n_1 n_2}$$

- $$\frac{d\sigma_\mu^2}{d\alpha} = 2\alpha\frac{1}{n_1}\sigma_1^2 - 2(1-\alpha)\frac{1}{n_2}\sigma_2^2 = 0$$

$$\alpha\left(\frac{1}{n_1}a\sigma_2^2 + \frac{1}{n_2}\sigma_2^2\right) = \frac{1}{n_2}\sigma_2^2$$

$$\alpha_o = \frac{n_1}{n_1 + an_2}$$

- if $a = 4$ and $n_1 = 2n_2$: $\alpha_o = \frac{2n_2}{2n_2+4.n_2} = 1/3$

$$\sigma_{\hat{\mu}}^2 n_1 n_2 = \sigma_2^2 n_2 (a.\alpha^2 + 2.(1-\alpha)^2)$$

For $\alpha = 1/3$, $\sigma_{\hat{\mu}}^2 = \frac{4}{3n_1}\sigma_2^2$.

For $\alpha = 1/2$, $\sigma_{\hat{\mu}}^2 = \frac{3}{2n_1}\sigma_2^2$.

So for $\alpha = 1/2$, the variance is 12.5 percent higher.

4 TD 4 :Maximum Likelihood Estimation

Prepare the 4 first exercices before the exercice session.

4.1 Poisson

Determine the maximum likelihood estimation of the parameter λ of a Poisson distribution, based on a random sample of size n.

$$f_X(x) = \frac{e^{-\lambda}\lambda^x}{x!}, \qquad x = 0, 1, \dots$$

Solution 4.1

$$L(\lambda) = f_X(x_1, x_2, ...x_n, \lambda) = \prod_i \frac{e^{-\lambda}\lambda_i^x}{x_i!} = e^{-n\lambda}\frac{\prod_i e^{x_i * \ln\lambda}}{x_i!}$$

$$\updownarrow(\lambda) = \ln L(\lambda) = -n\lambda + \sum_i x_i \ln\lambda - \sum_i \ln x_i!$$

$$\frac{dl(\lambda)}{d\lambda} = -n + \frac{1}{\lambda}\sum_i x_i = 0$$

$$\hat{\lambda} = \frac{\sum_i x_i}{n}$$

4.2 Delayed exponential

Let's have the exponential distribution, delayed by $\theta > 0$:

$$f_X(x) = \lambda e^{-\lambda(x-\theta)}, \qquad x \geq \theta$$

- Determine the ML estimators of λ and θ, based on a random sample of size n.
- describe a practical situation in which this model would make sense.

Solution 4.2

$$\ln L(\lambda,\theta) = n\ln\lambda - \lambda\sum_i (X_i-\theta), \qquad x_i \geq \theta$$

$$\frac{\partial \ln L(\lambda,\theta)}{\lambda} = \frac{n}{\lambda} - \sum_i (x_i-\theta) \Rightarrow \hat{\lambda} = \frac{1}{1/n\sum_{i=1}^n X_i - \theta)}$$

$$\frac{\partial \ln L(\lambda,\theta)}{\theta} = \lambda$$

So here it does not lead to any non trivial solution.
Hence, we have to come back to the delayed nature of the problem $x_i \geq \theta$.
Hence, the most likely solution for θ, especially if n is large, is $\theta = \min_i x_i$.
Note that this is a biased estimator, and an asymptotically unbiased estimator (like for the case of the lower/higher bound of a uniform ML estimation).
Alternatively, we can note the "conditinal" e likelihood function is

$$\ln L(\lambda,\theta|x_i) = n\ln\lambda - \lambda\sum_i (X_i-\theta), \qquad x_i \geq \theta$$

conditionned on the smallest value of $x_i \geq \theta$ (let's note it $x(1)$)
Now note that this is maximum

- iff $-\lambda\sum_i (X_i-\theta)$ is maximum subject to the restriction $x(1) \geq \theta$
- iff θ is maximum subject to the restriction $x(1) \geq \theta$
- iff $\theta = x(1)$.

So $x(1)$ is the ML estimate.

4.3 Geometric law

Let X be a geometric law of parameter p, find an ML estimator of p, based on a random sample of size n.

Solution 4.3

The likelihood function is given by:
$L(p) = (1-p)^{x_1-1}p(1-p)^{x_2-1}p...(1-p)^{x_n-1}p = p^n(1-p)^{\sum_1^n x_i - n}$
Taking log,
$\ln L(p) = n\ln p + (\sum_1^n x_i - n)\ln(1-p)$
Differentiating and equating to zero, we get,
$\frac{d[\ln L(p)]}{dp} = \frac{n}{p} - \frac{\left(\sum_1^n x_i - n\right)}{(1-p)} = 0$
Therefore,
$p = \frac{n}{\left(\sum_1^n x_i\right)}$
So, the maximum likelihood estimator of P is:
$P = \frac{n}{\left(\sum_1^n X_i\right)} = \frac{1}{\bar{X}}$
This agrees with the intuition because, in n observations of a geometric random variable, there are n successes in the $\sum_1^n X_i$ trials. Thus the estimate of p is the number of successes divided by the total number of trials.

4.4 Nameless law

Let X be a r.v. such that :

$$f_X(x) = \frac{1}{\theta^2} x e^{-x/\theta}, \qquad 0 \leq x < \infty, 0 < \theta < \infty$$

Determine the ML estimator of θ.

Solution 4.4

$$L(\theta) = \frac{1}{\theta^{2n}} \prod_i \left(x_i e^{-x_i/\theta} \right)$$

$$\ln L(\theta) = -2n \ln \theta + \sum_i \ln x_i - \sum_i \frac{x_i}{\theta}$$

$$\frac{d \ln L(\theta)}{d\theta} = -2n/\theta + \frac{\sum_i x_i}{\theta^2}$$

$$\hat{\mathbf{\Theta}} = \bar{X}/2$$

4.5 Uniform r.v.

Let X be a uniform r.v. on $[0, a]$.

- Show that the ML estimator of a is $\max X_i$ where $\max X_i$ is the largest estimation of X
- Intuitively, show that $\hat{a}_{ML}$ is biased.
- We can show that $\mathrm{E}[\hat{a}_{ML}] = a\frac{n}{n+1}$. Can we say that $\hat{a}$ consistently underestimates a ?
- Propose an unbiased estimator of a.

Solution 4.5

- $$L(a) = \prod_i \frac{1}{a} = \frac{1}{a^n}$$

  This likelihood function does not depend on X_i, so we have to choose $\hat{a}$ as the largest observation, which «maximizes» the likelihood function.
- the largest observation $X_i = x_{max}$ is smaller that a with probability one. So it does always underestimate a.
- if we take $\hat{a}_o = \frac{n+1}{n}\hat{a}_{ML}$, it becomes unbiased.

4.6 uniform r.v.

let X be a uniform r.v. $[0, a]$. We define $\hat{a}_1 = 2\bar{X}$ and $\hat{a}_2 = \dfrac{n+1}{n} \max X_i$, where $\max X_i$ is the largest observation of X in a random sample of size n. We can show that $\mathrm{var}[\hat{a_1}] = \dfrac{a^2}{3n}$ and that $\mathrm{var}[\hat{a_2}] = \dfrac{a^2}{n(n+2)}$. For $n > 1$, which is the best estimator ? Why ?

Solution 4.6

> Both estimators are unbiased, the second one has a smaller variance than the first, so it is obviously better (has a better MSE)

4.7 Raleigh distribution

The pdf of a Raleigh r.v. is given by :

$$f_X(x) = \frac{x}{\theta} e^{-x^2/2\theta}, \qquad x > 0, \qquad 0 < \theta < \infty$$

- Nowing that $\mathrm{E}\left[X^2\right] = 2\theta$, construct an unbiased estimator of θ.
- Determine the ML estimator of θ, compare with the previous one.
- Use the invariance property of the ML estimator to find the ML estimator of the median of a Raleigh r.v.

Solution 4.7

- $\mathrm{E}\big[X^2\big] = \sigma^2 + \mu^2 = 2\theta$

  Remember that $\mathrm{E}\big[S^2\big] = \sigma^2$, and $\mathrm{E}\big[\bar{X}^2\big] = \frac{\sigma^2}{n} + \mu^2$

  so we can take

$$\begin{aligned}
\hat{\mathbf{\Theta}}_U &= \frac{\frac{n-1}{n}S^2 + \bar{X}^2}{2} \\
&= \frac{\frac{1}{n}\sum_i (X_i - \bar{X})^2 + \bar{X}^2}{2} \\
&= \frac{\frac{1}{n}\sum_i (X_i)^2 - \frac{2}{n}\sum_i X_i\bar{X} + 2\bar{X}^2}{2} \\
&= \frac{\frac{1}{n}\sum_i (X_i)^2}{2}
\end{aligned}$$

- $$L(\theta) = \frac{1}{\theta^n}\prod_i x_i e^{-x_i^2/(2\theta)}$$

  $$\ln L(\theta) = -n\ln\theta + \sum_i \ln x_i - \sum_i \frac{x_i^2}{2\theta}$$

  $$\frac{d\ln L(\theta)}{d\theta} = -n/\theta + \frac{1}{2\theta^2}\sum_i x_i^2$$

  by setting the derivative to zero, we get :

  $$\hat{\mathbf{\Theta}}_{ML} = \frac{\sum_i X_i^2}{2n}$$

- they are identical,,
- The median m of the Raleigh function is given by :

  $$\int_0^m \frac{x}{\theta} e^{-x^2/2\theta} dx = \frac{1}{2}$$

  substituting $x^2/2\theta$ by t, this bowls down to $\int_0^m e^{-t}dt = \frac{1}{2}$

  Hence $\left[e^{-x^2/2\theta}\right]_0^m = -0.5$, leading to $m^2 = -2\theta\log(0.5) = 1.386\theta$.

  Then, using the invariance property, we have that

  $$\hat{m}_{ML} = 1.177\sqrt{\hat{\theta}_{ML}} = 1.177\sqrt{\frac{\sum_i X_i^2}{2n}}$$

4.8 Thickness of an oxide layer

Thichness measurements of an oxide layer on semiconductors are given by :

`425, 431, 416, 419, 421, 436, 418, 410, 431, 433, 423,`

```
426, 410, 435, 436, 428, 411, 426, 409, 437, 422, 428,
 413, 416.
```

- compute the point estimators of the mean and standard deviation
- compute the standard deviation of the mean estimator
- if the thickness of the oxide follows a normal law, give the ML estimates of μ and σ^2.
- Tracez la fonction de vraissemblance au voisinage de $\hat{\mu}$ et de $\hat{\sigma}^2$.

Solution 4.8

- Mean : $\bar{X} = 423.333$
- standard deviation $S^2 = (9.08256)^2$
- μ is equal to the sample mean, $\hat{\sigma}^2_{ML} = \frac{n-1}{n} S^2 = 79.06$

5 TD 5 : Estimation by Confidence Interval

5.1 Margarine

We measure the concentration of polyinsaturated fat in 6 blocks of margarine. The concentration measured are : 16.8 ; 17.2 ; 17.4 ; 16.9 ; 16.5 and 17.1. Give the confidence interval of the mean μ, based on these measurementent for a confidence lever of 99 %.

Solution 5.1

$\bar{x} = 16.9833; s = 0.3189.$
$n < 30$ and we don't know the population pdf, so we use a Student law with 5 degrees of freedom.
Here $\alpha = 0.01$ so $t_{0.005}(5) = 4.032$ and
$16.455 \leq \mu \leq 17.505$

5.2 Glass thickness

The thickness of glass plaque is controled. The mean of a sample of 25 plaques is measured as 4.05 mm and it's sample standard deviation is 0.08 mm. Give the one-sided confidence interval of the mean thickness, for a confidence interval of 95 % (we look for the following result : $\mu \geq \bar{x}_L$).

Solution 5.2

$n = 25; \bar{x} = 4.05mm, s = 0.08mm$
$t_{0.05}(24) = 1.711 \rightarrow \mu_L = 4.05 - 1.711 \times 0.08/\sqrt{25} = 4.023$

5.3 Quantity of beer

A controller of a brewery verifies the quantity of beer that the cafe holders put in a glass. On a sample of 25 glasses, he obtains a sample mean value ($\bar{x}$) of 24.5 cl with a sample standard deviation of 1.5 cl. Give the confidence interval of the quantity of beer served, at a confidence level of 95 %.

Knowing that a glass should contain 25 cl, can we pretend that the client is prejudiced ?

Solution 5.3

$t_{0.025}(24) = 2.064$, hence $23.8808 \leq \mu \leq 25.1192$ cl.
25 cl is in the confidence interval, so the client is not prejudiced "in the mean", but of course, if you look at the population, there is about 30 % of chance that a client gets less then 23 cl ... and is prejudiced. (about "in the mean" and "for my own realisation") ...

5.4 Targetting right

A laser has to target a given point. On a sample of 15 measures, we determine a standard deviation of $s = 0.008$ mm. Give the one-sided confidence interval of σ^2 at a confidence level of 99 %.

Solution 5.4

Here, we want to have a good accuracy, so we search $\mathbb{P}\big(\sigma^2 < s_\alpha^2\big) = 99\%$.
So :

- The parameter is σ^2.
- The statistic is S^2, and the normalized statistic is $X^2 = \dfrac{(n-1)S^2}{\sigma^2}$.
- The distribution of X^2 is a chi-squared distribution with $n-1$ degrees of freedom ($X^2 \sim \chi^2(\nu = n-1)$)
- Hence $\mathbb{P}\big(\sigma^2 \leq s_\alpha^2\big) = 99\%$ is equivalent to $\mathbb{P}\big(X^2 \geq \chi_\alpha^2\big) = 99\%$, where here in the table,
  - $\nu = 14$ (line number 14)
  - $\alpha = 0.01$

$\chi^2_{0.99}(14) = 4.66$ and $\sigma^2 \leq \left(\frac{(n-1)s^2}{\chi^2_{0.99}(14)}\right)$. This leads to $\sigma^2 \leq 0.192mm^2$ and $\sigma \leq 0.013mm$

5.5 Percentage of titanium in alloy

Percentage of titanium in alloy is measured on 51 parts. The sample standard deviation is $s = 0.37$. Give the two-sided confidence interval of σ at a confidence level of 95 %.

$0.31 < \sigma < 0.46$

5.6 Number of mobile phones

We want to know the number of people that have 2 mobile phones. What should be the sample size if we want to estimate this quantity with an error less than 2 % for a confidence level of 99 % ?

Solution 5.5

Considering the error of $\pm 2\%$, and the confidence level of 99 %, if the size is larger than 30, we have to use the normal distribution for the estimate of the proportion. $z_{\alpha/2} = 2.57$, and the error $2.57 \times \sqrt{\frac{p(1-p)}{n}} \leq 0.02$. This leads to $n \geq \frac{6.6p(1-p)}{0.02^2}$. If we take $p = 1/2$, this gives $p(1-p) = 1/4$, which is the largest value possible for all p. Hence : $n \geq \frac{6.6/4}{0.02^2} = 4125$.

5.7 Defects

An integrated circuits manufacturer counts 8 defect ICs on a sample of 1200 units. What is the confidence interval of the mean, at a confidence level of 95 %. Can we say that the number of defect ICs in the production line is less than 1%.

$0.0021 \leq \pi \leq 0.0113$, *so we can not pretend the defect rate is less than 1 %*

5.8 Yet Another Point Estimator

Two different measurement engines measure the voltage on n resistors. The voltage is a r.v. V. The two measurements are X_i and Y_i ($i = 1, \cdots n$).

We assume that the X_i and Y_i are independent and $X_i(Y_j)\, \mathcal{N}(\mu, \sigma^2)$.

- Show that the ML estimator of σ^2 is $\hat{\sigma}^2 = (1/4n) \sum_{i=1}^{N} (X_i - Y_i)^2$
- Show that $\hat{sigma}^2$ is a biased estimator of σ^2. Is it still biased for $n \to \infty$?
- Give an unbiased estimator of σ^2

6 TD 6 :Confidence intervals

6.1 Confidence interval

A die manufacturing machine yields 3 different line widths. A sample of 100 lines is manufactured for each width, yielding mean width of 0.09 mm, 1.2 mm and 2.6 mm. If the standard deviations are respectively 1 μm, 10 μm and 20 μm, determine the confidence intervals at levels 90 %, 95 % et 99 %.

Solution 6.1

For each case, we have here a bilateral confidence interval at, respectively, according to the normal table, at $\pm 1.65\sigma_{\bar{X}}, 1.96\sigma_{\bar{X}}$ and $2.58\sigma_{\bar{X}}$ ($F_Z(z)$ at .95, .975 and .995).
Taking into account that $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \sigma/10$
Hence, we have the following confidence intervals :

1.
   - $\mu \in [0.09 - 0.165 * 10^{-3}mm, 0.09 + 0.165 * 10^{-3}mm]$ with a confidence level of 90 %
   - $\mu \in [0.09 - 0.196 * 10^{-3}mm, 0.09 + 0.196 * 10^{-3}mm]$ with a confidence level of 95 %
   - $\mu \in [0.09 - 0.258 * 10^{-3}mm, 0.09 + 0.258 * 10^{-3}mm]$ with a confidence level of 99 %
2.
   - $\mu \in [1.2 - 1.65 * 10^{-3}mm, 1.2 + 1.65 * 10^{-3}mm]$ with a confidence level of 90 %
   - $\mu \in [1.2 - 1.96 * 10^{-3}mm, 1.2 + 1.96 * 10^{-3}mm]$ with a confidence level of 95 %
   - $\mu \in [1.2 - 2.58 * 10^{-3}mm, 1.2 + 2.58 * 10^{-3}mm]$ with a confidence level of 99 %
3.
   - $\mu \in [2.6 - 1.65 * 2 * 10^{-3}mm, 2.6 + 1.65 * 2 * 10^{-3}mm]$ with a confidence level of 90 %
   - $\mu \in [2.6 - 1.96 * 2 * 10^{-3}mm, 2.6 + 1.96 * 2 * 10^{-3}mm]$ with a confidence level of 95 %
   - $\mu \in [2.6 - 2.58 * 2 * 10^{-3}mm, 2.6 + 2.58 * 2 * 10^{-3}mm]$ with a confidence level of 99 %

6.2 Sample size

What must be the size of the sample such that, for the narrowest line, the precision would be $\pm 1\mu m$ at a confidence level of 95 %.

Solution 6.2

At a confidence level of 95 %, the precision is $\pm 1.96\sigma_{\bar{X}}$, so we need $\sigma_{\bar{X}} \simeq 0.5\mu m$, where $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{1\mu m}{\sqrt{n}}$ so here $n = 4$ would suffice.
BUT, $n = 4$ means small sample size, and we have to take a Student distribution rather than a Normal distribution (we did not assume X to be normal, By looking at the table, we find that $n = 7$ is ok, as then $t_{\alpha/2} = 2.447$ and $t_{\alpha/2}\sigma/\sqrt{7}$ is the first value smaller then 1.

6.3 TOEIC Scores

A TOEIC score before starting the studies at Polytech is performed for 91 students, with a mean score of 545. We admit that the sample is representative of the students entering in the Polytech network. What is the mean level (with a confidence interval of 90 % et 95 %), admitting a standard deviation of 75. What would happen if the standard deviation were 150 ?

Solution 6.3

The mean levels are $\bar{x} \pm 1.65\sigma_{\bar{X}}$ and $\bar{x} \pm 1.96\sigma_{\bar{X}}$, which gives confidence intervals of $[532, 558]$ and [529,560.5] for the first case ; $[519, 570]$ and $[514, 576]$ in the second case.
This means that a standard deviation of 150 gives a very wide range of scores ...

6.4 Revenue comparisons

A bank wants to determine the mean revenue of it's clients, and to determine if there is a difference between the mean revenue of men and women.

A sample of 225 men and a sample of 190 women is take, with a mean revenue of 22050 Euros with a standard deviation of 4000 Euros for men, and a mean revenue of 20300 Euros and a standard deviation of 3000 Euros for women. Compute the confidence intervals at a confidence level of 95 % and comment the difference in revenues.

What would be your conclusions if the standard deviation of both men and women's revenues was 6000 Euros ?

What would be your conclusions if the sample sizes were 30 men and 30 women, with the revenue statistics of the first part of this question ?

Solution 6.4

For men, $\sigma_{\bar{X}} = 4000/\sqrt{225} = 266.66 Euros$, hence the confidence interval is $22050 \pm 1.96 * 266.66$ = [21527,22572] Euros.
For women, $\sigma_{\bar{X}} = 3000/\sqrt{190} = 217.64 Euros$, hence the confidence interval is $20300 \pm 1.96 * 217.64$ = [19873,20726] Euros.
We conclude that the revenue of men is significantly higher that the revenue of women (even if here, "significantly" has not yet been defined).
If the standard deviation was 6000 Euros :
For men, $\sigma_{\bar{X}} = 6000/\sqrt{225} = 400 Euros$, hence the confidence interval is $22050 \pm 1.96 * 400$ = [21266,22834] Euros.
For women, $\sigma_{\bar{X}} = 6000/\sqrt{190} = 217.64 Euros$, hence the confidence interval is $20300 \pm 1.96 * 435$ = [19446,21153] Euros.
About the same conclusions, but, the highest revenue of women are very close to lowest revenue of men ....
For men, $\sigma_{\bar{X}} = 4000/\sqrt{30} = 730 Euros$, hence the confidence interval is $22050 \pm 1.96 * 730$ = [20618,23481] Euros.
For women, $\sigma_{\bar{X}} = 3000/\sqrt{30} = 548 Euros$, hence the confidence interval is $20300 \pm 1.96 * 548$ = [19226,21273] Euros.
Here, the two confidence intervals overlap, so we would conclude that the highest revenues of women is in the range of the lowest men's revenues ... due to "too small" samples.

6.5 Sample quality

A polling institute was contracted to make previsions on the results of the baccalaurÃ©at (more or less french equivalent to the chinese gaokao). A sample of 100 students of the scientific students yields a mean of 12.3 (on 20) with a standard deviation of 1 (we consider that this standard deviation is reflecting the real standard deviation of the population).

Give the conclusions of the polling institute.

Tell if the polling institute did a good job if, in reality :

- the mean result was 12.5 with a standard deviation of 3;
- the mean result was 12.4 with a standard deviation of 0.5 ;
- the mean result was 12.8 with a standard deviation of 5.

Solution 6.5

The "difficulty" here is that we work on the pdf of the original r.v., not on the pdf of the sample mean !
Here, we have $\bar{x} = 12.3$ and $s_{\bar{x}} = 0.1$.
Hence, $\sigma = 1$, so the conclusion of the polling institute would be that 99 % of the students would succeed ($z = -2.3 : \mathbb{P}(\text{success}) = 1 - F_Z{-}2.3$). 61.7 % would have a score above 12 (in France "mention assez bien"), and only 5 % would have a score above 14 (in France "Mention bien").
If in reality :

- the mean result was 12.5 with a standard deviation of 3, this would have lead to a success of only 82 %, so the prevision is not OK (the standard deviation is not OK).
- the mean result was 12.4 with a standard deviation of 0.5 ; the prevision on success rate is OK, but the prevision on number of scores above 12 would be 97.7 % (+2 σ)
- with a standard deviation of 5 : too large discrepancy in standard deviation

6.6 Sample size

The standard deviation of a r.v. over a population is $\sigma = 3$, what should the size of the sample be to estimate the mean μ with a error less than 0.5 at a confidence level of 95 %?

Solution 6.6

The error is ± 0.5 and we use a confidence interval of 95 %, this means that $0.5 = 1.96\sigma_{\bar{X}} = 1.96\sigma/\sqrt{n}$. Hence $n > (1.96 \times 3)^2 = 138.2$, and we take $n = 139$.

6.7 Manufacturing

The boss of a factory manufacturing robots wants to know the mean time a worker takes to assemble a robot.

- By observing 120 workers, he sees that the mean time to assemble one robot is 23 minutes. We suppose that the standard deviation is 4 minutes. Give the confidence interval of the mean time at 90 % and at 95 %.
- What should be the number of workers to be observed to have a precision of ± 15 seconds.

Solution 6.7

- $IC_{.95}$ = [22.28,23.71] minutes ; $IC_{.90}$=[22.40,23.60]
- at 95 %, this means 0.25 = $\frac{\sigma}{\sqrt{n}}$. Hence, $n = (1.96\sigma/0.25)^2 = 1024$

6.8 Efficiency of a server

To ensure the efficiency of a server, one must estimate the number of simultaneous users. According to measurements, for 100 different instants, the mean number of simultaneous users is 37.7, with a standard deviation of 9.2. Give the mean number of simultaneous users at a confidence interval of 90 %. Can we say that the number of simultaneous number is larger than 35 ? Discuss this question.

Solution 6.8

IC = [36.2 39.2]
The number is large with a confidence of 90 % at least. So let's let's compute the probability that the number is larger than 35, with is at 2.5 $\sigma_{\bar{X}}$ of the mean, hence, the probabiltiy that the mean number of simultaneous users is larger than 35 is 99.3 %.
But the probability that one particular number of simultaneous users is large than 35 is much lower. Indeed, consider that the real mean is 37.3, than 35 is only at 0.25 σ of the mean and the probability that the number of simultaneous users is larger than 35 is only 60 %.

7 TD 7 : Hypothesis testing

7.1 Expressions

For each of the expressions below, say if the expression is a correct hypothesis test :

1. $H_o : \mu = 25; H_1 : \mu \neq 25$ (Yes)
2. $H_o : \sigma > 10; H_1 : \sigma = 10$ (No, no equality in Ho)
3. $H_o : \bar{x} = 25; H_1 : \bar{x} \neq 25$ (No)
4. $H_o : \pi = 0.1; H_1 : \pi = 0.5$ (No, H1 equality)
5. $H_o : s = 25; H_1 : s > 25$ (No)

7.2 Quality control

A metallic cable manufacturer wants to verify the weight of it's cable holders. The manufacturer indicates that the holders have a weight of 100 kg with a standard deviation of 0.25 kg. The quality control engineer wants to test $H_o : \mu = 100kg$ against $H_1 : \mu < 100kg$, by using a sample of 4 holders.

- What is the error probability of type I, the critical region being defined by $\bar{x} = 99.5kg$?
- Find β if the true weight of the holders is 99.25 kg.
- Determine these quantities for $n = 16$.

Solution 7.1

$$\begin{cases} H_o : & \mu = 100 \text{ kg} \\ H_1 : & \mu < 100 \text{ kg} \end{cases}$$

for H_o true : $\bar{X} \sim T(100, 1/64)$
$\alpha = \mathbb{P}\big(\bar{X} \leq 99.5\big) = \mathbb{P}\Big(T_\alpha(3) \leq \frac{99.5-100}{1/8}\Big) = \mathbb{P}(T_\alpha(3) \leq -4) = \mathbb{P}(T_\alpha(3) \geq 4) \simeq .014$
$\beta = \mathbb{P}\Big(\hat{H}_o | H_1\Big) = \mathbb{P}\big(\bar{X} > 99.5 | \bar{X} \sim \mathcal{N}(99.25, 1/8)\big) = \mathbb{P}(T_\beta(3) > 0.25/(1/8)) = \mathbb{P}(T_\beta(3) > 2) \simeq 0.05$, and the power of the test is $1 - \beta \simeq 95\%$
for $n = 16$, for H_o true : $\bar{X} \sim \mathcal{N}(100, 1/16^2)$ $\alpha = \mathbb{P}\big(\bar{X} \leq 99.5\big) = \mathbb{P}\Big(T_\alpha(3) \leq \frac{99.5-100}{1/16}\Big) = \mathbb{P}(T_\alpha(3) \leq -8) = \mathbb{P}(T_\alpha(3) \geq 8) \simeq .002$
$\beta = \mathbb{P}\Big(\hat{H}_o | H_1\Big) = \mathbb{P}\big(\bar{X} > 99.5 | \bar{X} \sim \mathcal{N}(99.25, 1/16)\big) = \mathbb{P}(T_\beta(3) > 0.25/(1/16)) = \mathbb{P}(T_\beta(3) > 4) \simeq 0.025$, and the power of the test is $1 - \beta \simeq 97.5\%$

7.3 Capacitor testing

A manufacturer tests a new method to manufacture capacitors. Express in μF, the capacitance is approximately normally distributed.

The manufacturer test the hypothesis $H_o : \mu = 175\mu F$ against $H_1 : \mu > 175\mu F$ with a sample of $n = 10$ capacitors. The standard deviation of the sample is 20 μF.

- Find α if the critical region is $\bar{x} > 185\mu F$
- What is the error probability of type II if the true value of the mean is 195 μF ?
- Let $\bar{x} = 190\mu F$, what are the conclusions of the test ?
- If the true value is $\mu = 175\mu F$, and $\bar{x} = 190\mu F$, is this value really unusual ? (base your conclusions on $\mathbb{P}(\bar{x} > 190\mu F)$)

Solution 7.2

$$\begin{cases} H_o: & \mu = 175\mu F \\ H_1: & \mu > 175\mu F \end{cases}$$

for H_o true : $\bar{X} \sim \mathcal{N}(175, \frac{20}{\sqrt{10}})$

- $\alpha = \mathbb{P}\big(\bar{X} > 185\big) = \mathbb{P}\big(T_\alpha(9) > \frac{185-175}{6.325}\big) = \mathbb{P}(T_\alpha(9) > 1.58) \simeq 0.075$
- $\beta = \mathbb{P}\Big(\hat{H}_o | H_1\Big) = \mathbb{P}\big(\bar{X} < 185 | \bar{X} \sim \mathcal{N}(195, 6.325)\big) = \mathbb{P}(T_\beta(9) < -10/6.325) = \mathbb{P}(T_\beta(9) > 1.58) \simeq 0.075$, and the power of the test is $1 - \beta \simeq 92.5\%$

  Here, the critical region is exactly in the middle of the two means (of H_o and H_1, so $\alpha = \beta$).
- If $\bar{x} = 190\mu F$, then $\mathbb{P}(\bar{x} > 190\mu F) = \mathbb{P}(T_\alpha(9) > 15/6.225) = 0.025$, so the test is OK and the significance of the test is between "significant" and "very significant".

7.4 Alternated traffic

A sample of 500 Paris citizen is polled to know if they are favorable to "alternated traffic" in case of atmospheric pollution. If more then 315 people are in favor of alternated traffic, we conclude that at least 60 % of the voters are in favor.

- What is the type I error if exactly 60 % of the Paris citizen are in favor of alternated traffic ?
- What is the type II error (β) if 75 % of the citizen are in favor of alternated traffic ?

Suggestion : use the approximation of a binomial r.v. by a normal r.v.

Solution 7.3

- $n = 500$, $X \sim Bi(n, p)$ with $\mu_X = n.p$ and $\text{var}[X] = np(1-p)$.

  Hence, we approximate $X \sim \mathcal{N}(n/2, np(1-p))(\mathcal{N}(250, 120))$ (under hypothesis $H_o : \pi = 0.6$)

  Then, $\alpha = \mathbb{P}\left(Z \geq \frac{X-np}{\sqrt{np(1-p)}}\right) = \mathbb{P}(Z \geq 1.369) = 0.0853$

  (Note that we can also work on proportions, in which case we have $\alpha = \mathbb{P}\left(Z \geq \frac{(X-np)/n}{\sqrt{p(1-p)/\sqrt{n}}}\right)$ which is exactly the same.)
- $\beta = \mathbb{P}\left(Z \leq \frac{X-np}{\sqrt{np(1-p)}}\right) = \mathbb{P}\big(Z \leq (315-375)/\sqrt{93.75}\big) = \mathbb{P}(Z \leq 6.2) \simeq 0$

7.5 Olive oil production

We study the olive production, assumed to be normal (Gaussian) with $\sigma = 3kg$. The production of 5 olive trees is : 91.6 kg, 88.75 kg, 90.8 kg, 89.95 kg et 91.3 kg.

- Can we say that the mean production of an olive tree is 90 kg ?
- What is the p-value of this test ? with a probability of 0.95 ?
- What is the power of the test $(1-\beta)$ if we fix $H_1 : \mu = 92kg$.

Solution 7.4

$n = 5, \bar{x} = 90.48, s^2 = 1.3275 s = 1.1514$

$$\begin{cases} H_o : & \mu = 90kg \\ H_1 : & \mu \neq 90kg \end{cases}$$

The statistic is $\bar{X}$ and the distribution is normal, because σ is known and X is normally distributed.

- Taking a significance level at 5 %, we have $z_{\alpha/2} = 1.96$ and $\bar{x}_c = \bar{x} \pm 1.65 * 3/\sqrt{5} = [87.8, 93.11]$, hence we can say that the mean is 90 kg.
- The p-value is based on $2.\mathbb{P}(X > 90.48) = \mathbb{P}(Z > (0.48/(1.3416)) = 7Z\%$ (here $p - value$ has no real sense ...)
- $\beta = \mathbb{P}\Big(\hat{H}_0 | H_1\Big)$. Having $\bar{x}_H \simeq 92$, we have $\beta = \mathbb{P}\big(\bar{X} > 87.8 | \mu = 92\big) + \mathbb{P}\big(\bar{X} < 93.1 | \mu = 92\big) = 0.8$, and the power is 1-0.8=0.2

7.6 Compare two production lines

A constructor wants to compare an old production line with a new experimental one.

Defect parts

He first make a qualitative control. The first line has 12 defect parts for 88 good parts. The experimental line has 20 defect parts for 122 good parts.

1. Compute the confidence interval (at 95 %) of the proportion of bad parts in each line.
2. Compute the confidence interval (at 95 %) of the difference of proportions of bad parts in each line.
3. Can we say that the new line is better ? Express the hypotheses of the test and the decision rules. Compute the p-value.

Life duration

The second test is quantitative. We sample the two lines and measure the lifetime expressed in days. The number of individuals is small as it is a destructive tes. We assume the lifetime has a normal distribution.

The first sample yields the following values:
101.0 103.0 103.0 88.2 108.0 102.0 100.0 93.5 96.4 94.8
The second sample yields the following values:
118.8 116.0 112.7 102.3 115.0 106.3 107.6.

1. Can we say that the new line is better ? Express the hypotheses of the test and the decision rules. Compute the p-value.

2. If the values for the second line were :
   120.3 117.0 113.0 100.6 115.8 105.3 106.8
   would the test procedure change ?

Solution 7.5

$n_1 = 100, n_2 = 142, \hat{p}_1 = 12/100 \hat{p}_2 = 20/142$
For line 1 : $IC = [\hat{p}_1 \pm 1.96\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{100}} = [0.056, 0.184]$
For line 2 : $IC = [\hat{p}_2 \pm 1.96\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} = [0.0836, 0.198]$
For the difference $IC = (\hat{p}_1 - \hat{p}_2) \pm 1.96.\sqrt{\frac{\hat{p}_2(1-\hat{p}_2)}{n_2} + \frac{\hat{p}_1(1-\hat{p}_1)}{n_1}} = [-0.106, 0.064]$
We can not say the new one is better (nor worse) as the hypothesis H_o : $\pi_1 = \pi_2$ is accepted.
For the samples given :
$n_1 = 10, \bar{x}_1 = 98.99, s_1^2 = 33.14$
$n_2 = 7, \bar{x}_2 = 111.24, s_2^2 = 35.61$
We first have to decide if $\sigma_1 = \sigma_2$, hence :

1. $H_o : \sigma_1 = \sigma_2; H_1 : \sigma_1 \neq \sigma_2$
2. $\alpha = 0.05$
3. Statsitic : s_1^2/s_2^2
   Distribution : $F = \frac{S_1^2}{S_2^2}$
4. critical values and decision region :
   $f_{.975}(9,6) < f_o < f_{.025}(9,6)$ i.e. $1/1.61 < f_o < 5.52$. because $f_{1-\alpha}(\nu_1, \nu_2) = f_\alpha(\nu_2, \nu_1)$.
5.
6. Here $f_o = 33.14/35.61 = 0.93$, so we accept H_o.

Now we can do the test for the mean :

1. $H_o : \mu_1 = \mu 2; H_1 : \mu_2 > \mu_1$
2. $\alpha = 5\%$
3. Statisctic $x_1 - x_2$
   Distribution : we have $\sigma_1 = \sigma_2$ hence, we have a Student distribution with

$$T = \frac{(\overline{X}_2 - \overline{X}_1) - (\mu_2 - \mu_1)}{S_c\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

   $\nu = n_1 + n_2 - 1$, and $S_c^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{(n_1-1)+(n_2-1)}$
   So here $\nu = 15$ and $S_c^2 = 34.13$
4. The critical value is $t_{0.05}(15) = 1.753$ and we accept H_o if $t_o < t_{0.05}(15)$.
5. $t_o = (111.24 - 98.99)/(\sqrt{34.13} * 0.492) = 4.26$, and we reject H_o

7.7 Automatic filling

To fill bottles with liquid, we use an automatic filling machine. The manufacturer of this machine indicates that the filling volume has a normal distribution. On a sample of 20 bottles, the sample variance is 0.0153.

- If the variance of the filling is larger then 0.01, an too large proportion of bottles will be over or under-filled. Are there proofs in the data from the samples suggesting that the manufacturer has a problem with bottles being under or over filled ? Use $\alpha = 5\%$
- What is the p-value of this test ?

Solution 7.6

$n = 20, s^2 = 0.0153, \alpha = 0.05$

$$\begin{cases} H_o: & \sigma^2 = 0.01 (\sigma_o) \\ H_1: & \sigma^2 < 0.01 \end{cases}$$

Statistic : $\chi^2 = \frac{(n-1)S^2}{\sigma_o^2}$
Critic values $\chi_{1-\alpha}(19) = 10.12$
Sample value $\chi^2 = \frac{(n-1)s^2}{\sigma_o^2} = 19 * 0.0153/0.01 = 29.07$
And $10.12 < 29.07$ so we accept H_o.
The p-value is $1 - \mathbb{P}\big(X^2 > 29.07\big) \simeq 0.95$.

7.8 Rivets in a hole

A rivet must be inserted in a hole. A random sample of 15 parts is selected ant the diameter of the hole is measured. The standard deviation of the measurements is 0.008 mm. Give the 99 % IC for the variance.

There is a high chance that the rivet is not suited to the hole if the standard deviation of the diameter is larger than 0.01 mm.

1. Can we say that the standard deviation is larger than 0.01 mm ?
2. What is the p-value of the test ?
3. Redo with $\alpha = 0.05$

Solution 7.7

The problem suggests a one-sided IC and test, with $\chi^2_{\alpha/2}(n-1) = \chi^2_{0.005}(n-1) = 31.32$ $\chi^2_{1-\alpha/2}(n-1) = \chi^2_{0.995}(n-1) = 4.07$
Hence, the $\sigma_H = \sqrt{\frac{(n-1)s^2}{\chi^2_{1-\alpha/2}(n-1)}} = 0.014mm$, $\sigma_L = \sqrt{\frac{(n-1)s^2}{\chi^2_{\alpha/2}(n-1)}} = 0.00534mm$, and IC =$(0.00534mm)^2, (0.014mm)^2]$

$$\begin{cases} H_o: & \sigma^2 = 0.01^2 (\sigma_o) \\ H_1: & \sigma^2 < 0.01^2 \end{cases}$$

This is a one-sided test, and we accept H_o if $\chi^2_o < \chi^2_{1-\alpha}(n-1) = \chi^2_{0.99}(n-1) = 4.66$, here $\chi^2_o = 8.96$ so we accept H_o (meaning that we decide that the precision of the hole is not good).
The p-value is $\mathbb{P}\big(X^2 < 8.96\big)$ is between .1 and .5, so we have a high p-value.

7.9 Comparison of two variances

An engineer wants to test the precision of a new device (B) compared to device (A). He makes respectively 13 and 15 tests on A and B, and observes variances of 6.3 and 3.2.

Express a on-sided hypothesis test.

What is the conclusion of this test for a confidence level at 90 %.

Solution 7.8

$n_A = 13, n_B = 15, s_A^2 = 6.3, s_B^2 = 3.2$

$$\begin{cases} H_o: & \sigma_A^2 = \sigma_B^2 \\ H_1: & \sigma_A^2 > \sigma_B^2 \end{cases}$$

$f_o = \frac{s_A^2}{s_B^2} = 1.97$

We will test if $f_o < f_\alpha(n_A - 1, n_B - 1) = 2.05$, which is the case, so we accept H_o.

7.10 Thickness of substrate layer

On a substrate, the thickness of a plastic layer is influence by the application temperature. In a random experiment, the layers of 11 substrates are realised at a temperature of 125 degrees, yielding a mean thickness of 103.5 with a standard deviation of 10.2. On another sample of 13 substrates, at a temperature of 150 degrees, the sample mean is 99.7 with a standard deviation of 20.1.

- considering the variances of the two samples are different, for $\alpha = 0.01$, can we say that the increase in temperature reduces the thickness of the layer ?
- what is the p-value of the test ?
- Test the variances : what can you conclude ?

For $\mu_1 - \mu_2$:

$$\begin{cases} H_o: & \mu_1 = \mu_2 \\ H_1: & \mu_1 > \mu_2 \end{cases}$$

with $n_1 = 11, n_2 = 13, s_1 = 10.2, s_2 = 20.1, \hat{x}_1 = 103.5, \hat{x}_2 = 99.7$

The critical value is $\bar{\Delta x}_c = (t_\alpha(\nu)\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, where, according to the formulas given, ν=18

which gives $\bar{\Delta x}_c = 2.878 * 6.367 = 18.32$ and we accept H_o.

Here the p-value is between 25 and 68 %

For the variances

$$\begin{cases} H_o: & \sigma_1^2 = \sigma_2^2 \\ H_1: & \sigma_1^2 < \sigma_2^2 \end{cases}$$

Whe have f_o = 0.2575,

$f_{1-\alpha}(10, 12) = 1/f_\alpha(12, 10) = 1/4.71 = 0.212$

Hence, $f_o > f_{1-\alpha}(10, 12)$ and we accept H_o.

Note that at 5 %, $f_{1-\alpha}(10, 12) = 0.34$, and we would have rejected H_o

7.11 Influence du Cadmium sur le taux de glucose

A physiologist studies the influence of cadmium on the amount of glucose in blood. He fills 2 basins, one with pure water, the other with adding 0.01 mg of Cd per liter. 18 trouts are put in each of the basis, and the amount of glucose in blood is measured after 2 hours.

The results are that, in the pure water basis, the mean quantity is 86.6 with a variance of 5.1 and in the Cd added basin, a sample mean of 91.2 with a variance of 8.3.

The expected variances were 5 in the pure water and 10 in the Cd added water.

Can we say that adding cadmium increases the amount of sugar for the trouts ?

We consider index 1 : pure water, and 9 trouts in each basin.

1. Ho: $\mu_1 = \mu 2$
   H1: $\mu_1 > \mu_2 \qquad (\mu_2 - \mu_1 < 0)$ (H_o corresponds to the case where cadmium increases sugar in blood)
2. $\alpha = 1$ et $\alpha = 5\%$
3. $n_1 = n_2$, Hence Student with $n_1 + n_2 - 2$ d.o.f. S_c^2 = 7.5
4.
5. Decision rule:
   - Reject Ho if $t < t_\alpha$
   - Reject if $\bar{x_2} - \bar{x_1} < t_\alpha . Sc. \sqrt{1/n_1 + 1/n_2}$ (H1: $\mu_2 - \mu_1 < d_o$ and we test $\bar{x_2} - \bar{x_1}$!).
6. Calculs
   - $t = \frac{91.2 - 86.6}{Sc.\sqrt{2/9}} = 3.563$;
   - $t_{1\%} = cdft("T", 16, 0.99, 0.01) = 2.583$; $t_{5\%} = cdft("T", 16, 0.99, 0.01) = 1.745$
   - $t > t_\alpha$: accept Ho
   - $\bar{x_2} - \bar{x_1}$ (= 4.6) > $t_\alpha . Sc. \sqrt{1/n_1 + 1/n_2}$ (= 3.33 for 1%)
7. Hence Cadmium increases glycemia significantly.

8 TD 8 : Chi-2 tests

8.1 Testing a law

Let X be a random variable, whose observation table is given by
$o(0) = 24, o(1) = 30, o(2) = 31, o(3) = 11, o(4) = 4$.

Based on these observations, can we say that X follows a Poisson distribution with mean 1.2, with a confidence level of 95 % ? What is the p-value of this test.

Solution 8.1

Let's first compute the expected values. The expected values are $e(i) = \mathbb{P}(X = i) = \frac{1.2^i e^{-1.2}}{i!}$, which gives (given $n = 100$) $e = [30.12, 36.14, 21.69, 8.672.6]$
Which leads to a value $\chi^2_o = 7.66$.
The distribution of the statistic is a chi-2 with (5-1) degrees of freedom (the parameter used is not taken from the statistics).
Hence the hypothesis testing is :

$$\begin{cases} H_o : & \text{the r.v. is not Poisson distributed} \\ H_1 : & \text{the r.v. is Poisson distributed} \end{cases}$$

And the critical value is $\chi^2_{1-\alpha}(4) = 9.48$ and as $\chi^2_o = 7.66 < \chi^2_{1-\alpha}(4)$ we accept H_o, and hence the r.v. is potentially Poisson distributed
The p-value is approximately 10 %.

8.2 Normality test

The japanese Sumo association wants to verify the following hypothesis : the weight of Sumos follow a normal law of mean $\mu = 145$ kg and a standard deviation $\sigma = 10$ kg.

The association weights a sample of

Poids	FrÃ©quence
< 125 kg	6
125 $\leq$ 135 kg	36
135 $\leq$ 145 kg	113
145 $\leq$ 155 kg	93
155 $\leq$ 165 kg	39
> 165 kg	13

Can we conclude that the normality assumption is fullfilled.

Solution 8.2

Based on the mean and variance, the probabilities in each classes can be computed as :

$$[0.022750.135910.341340.341340.135910.02275]$$

Which means that in each class, the expected number of individuals is given by :

$$[6.8250440.77154102.40342102.4034240.771546.82504]$$

Hence, the χ^2 value is 8.282, and the number of degrees of freedom is $\nu = 6 - 1$.
Then, the p-value is $\mathbb{P}X^2(5) > 8.282 = 0.14$, and we can not reject Ho, hence we conclude that the measurements are compatible with the Sumo's association hypothesis.

8.3 Independence test

Taken from Louis Houde, UniversitÃ© du QuÃ©bec Ã Trois-RiviÃ¨res

To have a better understanding of client's interes in a new product, a company poll 79 persons (who answer). The interest in the product is "no interest" or "minor interest" or "great interest". The family situation is also taken into account (has at least one child : Yes/No). We want to verify if the interest in the product is linked to the family situation. The results are the following :

Child/Interest	no	minor	great
yes	10	12	3
no	7	38	9

Solution 8.3

The Ho hypothesis is that X and Y are independent.
The contingency table with theoretical frequencies is then given by :

Child/Interest	no	minor	great
yes	10 (5.4)	12 (15.8)	3 (3.8)
no	7 (11.6)	38 (34.2)	9 (8.2)

Conditions are OK.
Leading to a chi2 value of 7.401
The number of degrees of freedom is (2-1).(3-1)=2
Hence the p-values is $\mathbb{P}X^2(2) > 7.401 = 0.024$ et on refuse Ho Ã un niveau de 5 %, on accepte Ho Ã un niveau de 1%.

8.4 Covid-19 Questions

8.4.1 Link between smoking and Covid-19

This question is based on the following study :

"*Clinical Characteristics of Coronavirus Disease 2019 in China*",

Guan, Wei-jie and Ni, Zheng-yi and Hu, Yu and Liang, Wen-hua and Ou, Chun-quan and He, Jian-xing and Liu, Lei and Shan, Hong and Lei, Chun-liang and Hui, David S.C. and Du, Bin and Li, Lan-juan and Zeng, Guang and Yuen, Kwok-Yung and Chen, Ru-chong and Tang, Chun-li and Wang, Tao and Chen, Ping-yan and Xiang, Jie and Li, Shi-yue and Wang, Jin-lin and Liang, Zi-jing and Peng, Yi-xiang and Wei, Li and Liu, Yong and Hu, Ya-hua and Peng, Peng and Wang, Jian-ming and Liu, Ji-yang and Chen, Zhong and Li, Gang and Zheng, Zhi-jian and Qiu, Shao-qin and Luo, Jie and Ye, Chang-jiang and Zhu, Shao-yong and Zhong, Nan-shan, **New England Journal of Medicine**, February 2020, URL : `https://doi.org/10.1056/NEJMoa2002032`

According to this table, can we say that Smoking makes you better immune against Covid-19 ? We assume that the proportion of smokers in China is 27.7 % (Parascandola M, Xiao L. *Tobacco and the lung cancer epidemic in China.* Transl Lung Cancer Res. 2019;8(Suppl .iso 1):S21-S30. doi:10.21037/tlcr.2019.03.12).

Solution 8.4

8.4 → What do we have in the data

(1) Sample of a population of ill persons

on smoking

(2) persons that smoked or not

what is the question ?

⇒ Can we say that smoking makes you better immune against Covid 19

in terms of probability

→ $P(Ill | Smoker) < P(Ill | \overline{Smoker})$

from the table we can compute

$P(Smoker | Ill)$ → write this as $\hat{p}$

$P(Ill | Smoker) = P(I|S) = P(S|I) \cdot \frac{P(I)}{P(S)}$

$P(I|\bar{S}) = P(\bar{S}|I) \cdot \frac{P(I)}{P(\bar{S})}$

$P(I|S) < P(I|\bar{S})$

$\Leftrightarrow P(S|I) < P(\bar{S}|I) \cdot \frac{P(S)}{P(\bar{S})} \Rightarrow$

$P(S|I) < P(\bar{S}|I) \times \frac{0.277}{0.723}$

$< P(\bar{S}|I) \cdot 0.383$

$P(\bar{S}|I) = (1 - P(S|I)) \cdot 0.383 < 0$

$P(S|I) \times 1.383 < 0.383$

$P(S|I) < 0.277$

Looks categorically correct!

→ $H_0 : \pi = 0.277$

$H_1 : \pi < 0.277$

from data $n = 1085$

$\hat{p} = \frac{153}{1085} = 0.141$

$H_0 : \hat{P} \sim N(\pi, \frac{\pi(1-\pi)}{n})$ because $n \gg 30$

$n\hat{p} > 5$

$n(1-\hat{p}) > 5$

$\hat{p} \neq 0 \quad \hat{p} \neq 1$

$\sigma^2_{\hat{p}} = 0.00011$

$\sigma_{\hat{p}} = 0.0105$

$\alpha_{p/2} = cdf\left(\frac{\hat{p} - \pi}{\sigma_{\hat{p}}}\right) = cdf(-12.87) = 0$

⇒ We very strongly reject H_0

⇒ Smoking definitely protects against Covid-19

beware → p-value extremely small does not tells that it absolutely protects → it would be another Hypothesis ($H_0 : \pi = 0$)

8.4.2 Law of number of deaths according to the age

Vous pouvez trouvez ici : `https://www.thelancet.com/pdfs/journals/laninf/PIIS1473-3099(20)30243-7.pdf` l'article : *Estimates of the severity of coronavirus disease 2019: a model-based analysis*, par Robert Verity*, Lucy C Okell*, Ilaria Dorigatti*, Peter Winskill*, Charles Whittaker*, Natsuko Imai, Gina Cuomo-Dannenburg, Hayley Thompson, Patrick G T Walker, Han Fu, Amy Dighe, Jamie T Griffin, Marc Baguelin, Sangeeta Bhatia, Adhiratha Boonyasiri, Anne Cori, Zulma CucunubÃ¡, Rich FitzJohn, Katy Gaythorpe, Will Green, Arran Hamlet, Wes Hinsley, Daniel Laydon, Gemma Nedjati-Gilani, Steven Riley, Sabine van Elsland, Erik Volz, Haowei Wang, Yuanrong Wang, Xiaoyue Xi, Christl A Donnelly, Azra C Ghani, Neil M Ferguson*.

At page 5 of this paper, the table could indicate that the proportion of the number of deaths due do Covid-19 follows an exponential law as a function of age. Examine this hypothesis with the data given in the paper. Also examine this hypothesis for the proportion of hospitalised persons (table 3).

Solution 8.5

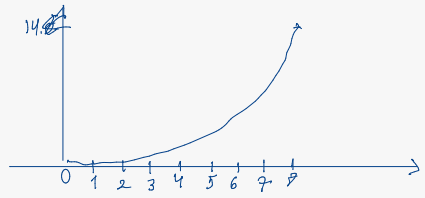
8.3 · data

Class	Age	death	cases	Crude %	Expected
0	0-9	0	416	0.	0
1	10-19	1	549	0.182	0.049
2	20-29	7	3619	0.193	0.18
3	30-39	18	7600	0.237	0.48
4	40-49	38	8571	0.443	0.99
5	50-59	130	10.008	1.30	1.93
6	60-69	309	8583	3.60	3.72
7	70-79	312	3918	7.96	7.26
8	≥80	208	1408	14.8	14.8
	total	1023	44672	2.29%	

here x is the class number

looks like → fit the curve

$P \simeq e^{-\lambda x} - 1$



⇒ for example take

$\lambda = \frac{\ln(14.8+1)}{8}$

$= 0.0431$

Δ! O_i = Ratio · N total

E_i = Ratio · N

⇒ chi 2 = 508.14

dof = 9 − 1 − 1 = 7

⇒ p-value = 0

⇒ the model is not correct.

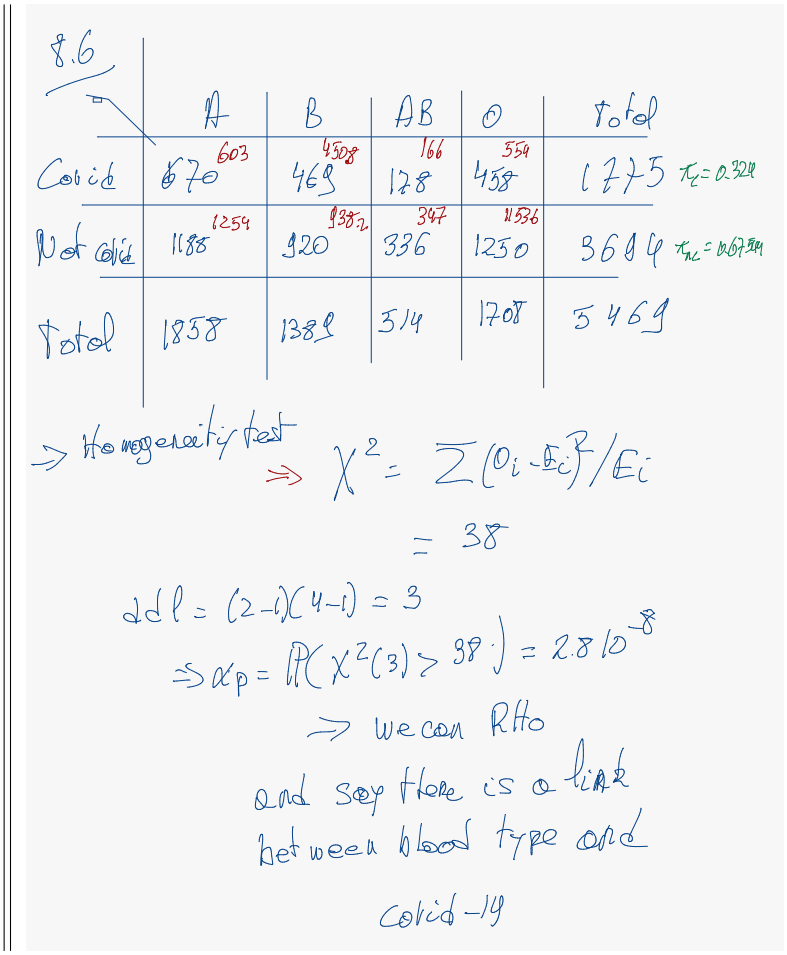
8.4.3 Relation entre groupe sanguin et Covid-19

Here is the paer on which we base this exercice : `https://www.medrxiv.org/content/10.1101/2020.03.11.20031096v2` l'article *Relationship between the ABO Blood Group and the COVID-19 Susceptibility (article)* by : Zhao, Jiao and Yang, Yan and Huang, Hanping and Li, Dong and Gu, Dongfeng and Lu, Xiangfeng and Zhang, Zheng and Liu, Lei and Liu, Ting and Liu, Yukun and He, Yunjiao and Sun, Bin and Wei, Meilan and Yang, Guangyu and Wang, Xinghuan and Zhang, Li and Zhou, Xiaoyang and Xing, Mingzhao and Wang, Peng George.

This paper suggests there is a link between blood type and Covid-19.

According to the data, build a test that proves this.

Solution 8.6



9 Linear Regression

9.1 A full linear regression problem

An article in Concrete Research ("Near Surface Characteristics of Concrete: Intrinsic Permeability," Vol. 41,1989), presented data on compressive strength X and intrinsic permeability y of various concrete mixes and cures.

Summary quantities are $n = 14, \sum y_i = 572, \sum y_i^2 = 23530, \sum x_i = 43, \sum x_i^2 = 157.42$,and $\sum x_i y_i = 1697.80$. Assume that the two variables are related according to the simple linear regression model.

- Calculate the least squares estimates of the slope and intercept.
- Use the equation of the fitted line to predict what permeability would be observed when the compressive strength is $x = 4.3$.
- Give a point estimate of the mean permeability when compressive strength is $x = 3.7$.
- Suppose that the observed value of permeability at $x = 3.7$ is $y = 46.1$. Calculate the value of the corresponding residual.
- Estimate σ^2 and the standard deviation of $\hat{\beta}_1$. (Note that $\sum (y_i - \hat{y}_i)^2 = \sum y_i^2 - n\bar{y}^2 - \hat{\beta}_1 S_{xy}$.
- Test for significance of regression using $\alpha = 0.05$. Find the P-value for this test. Can you conclude that the model specifies a useful linear relationship between these two variables?
- What is the standard error of the intercept point in this model?
- Find a 95% confidence interval on each of the following:
  - Slope
  - Interception point
  - Mean permeability when $x = 2.5$
  - Find a 95% prediction interval on permeability when $x = 2.5$. Explain why this interval is wider than the previous interval.

Solution 9.1

1.1 + $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

$S_{xx} = 157.42 - \frac{43^2}{14} = 25.3428571$

$S_{xy} = 1697.8 - \frac{43 \times 572}{14} = -59.057$

$\hat{\beta_1} = \frac{S_{xy}}{S_{xx}} = -2.330$

$\hat{\beta_0} = \bar{y} - \hat{\beta_1}\bar{x} = \frac{572}{14} - (-2.330)\frac{43}{14} = 48.013$

+ $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$

$x = 4.3 \Rightarrow \hat{y} = 37.99$

$x = 3.7 \Rightarrow \hat{y} = 39.39$

+ $e = y - \hat{y} = 46.1 - 37.99 = 8.71$

+ $H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

$\Rightarrow \alpha = 0.05$

$\Rightarrow t = \frac{\hat{\beta_1}}{\Delta\hat{\beta_1}}$

$\Rightarrow$ Rh0 if $t_0 < t_{0.025, 12}$ ou $t_0 > t_{0.025, 12} = 2.179$

$\Rightarrow t_0 = \frac{-2.330}{\Delta\hat{\beta_1}} = -8.64$ $\Rightarrow$ p-val $= 1.7 \cdot 10^{-6}$

$\Delta\hat{\beta_1} = \frac{\hat{\sigma}}{\sqrt{S_{xx}}}$ | $\hat{\sigma}^2 = \frac{1}{n-2}\left[\sum y_i^2 - n\bar{y}^2 - \hat{\beta_1} S_{xy}\right]$

$= 1.844$

$= 0.27$

* IC on β_1 : $\hat{\beta_1} \pm 2.179 \times 0.27$ | $-2.9175 < \beta_1 < -1.7425$

β_0 : $\hat{\beta_0} \pm 2.179 \times 0.596$ | $46.715 < \beta_0 < 49.3114$

on μ for $x_0 = 2.5$

$\hat{\mu} = 48.013 - 2.33(2.5) = 42.189$

IC : $42.189 \pm 2.179 \sqrt{1.844\left(\frac{1}{14} + \frac{(2.5 - 3.071)^2}{25.349}\right)}$

0.399

statap_2ndsem_tds 1 $41.33 < \hat{\mu}_{x_0} < 43.048$ 1 sur 1

IC on $\hat{y}$

$\Rightarrow \hat{y_0} \pm t_{n/2} \sqrt{\hat{\sigma}^2\left(1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{S_{xx}}\right)}$

$38.23 < y_0 < 46.1281$