Exercises in Statistical Inference

(A collection of exams in TAMS24)

Johan Thim, MAI

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk statistik

Kurskod: TAMS24
Provkod: TEN1
Datum: 2018-10-22
Institution: MAI

Exam in Statistics

TAMS24/TEN1 2018-10-22 8–12

You are permitted to bring:

- a calculator (no computer);
- Formel- och tabellsamling i matematisk statistik (from MAI);
- Formel- och tabellsamling i matematisk statistik, TAMS65;
- TAMS24: Notations and Formulas (by Xiangfeng Yang).

Grading: 8-11 points giving grade 3; 11.5-14.5 points giving grade 4; 15-18 points giving grade 5. Your solutions need to be complete, well motivated, carefully written and concluded by a clear answer. Be careful to show what is random and what is not. Assumptions you make need to be explicit. The exercises are in number order.

Solutions can be found on the homepage a couple of hours after the finished exam.

1. Let $f(x\,;\theta)$ be the density function for the Gamma distribution (with unknown parameter θ),

$$f(x\,;\theta) = \frac{\theta(\theta x)^{v-1}e^{-\theta x}}{\Gamma(v)},$$

where $x > 0$ and $v > 0$.

   (a) Find the ML-estimate $\widehat{\theta}$ for θ. (2p)

   (b) Show that the estimate $1/\widehat{\theta}$ is an unbiased estimate of $1/\theta$. *Hint: If X is Gamma distributed like above, then $E(X) = v/\theta$.* (1p)

2. During one hundred days, Lena and Sture has been collecting used syringes at a central cemetery that's used by the local drug addicts. They've written down the number found each day and the corresponding frequencies can be found below.

Number found	0	1	2	3	4
Frequency	38	33	26	2	1

   Lena believes the data is Po(1)-distributed (Poisson distributed with expectation 1), but Sture disagrees. Use a suitable test to see if Sture is correct in rejecting the hypothesis at the 1%-level. (2p)

3. Belinda is a hobby chemist experimenting with organic peroxides. She's trying to synthesize hexamethylene triperoxide diamine (HMTD) using two slightly different methods. Method one uses citric acid and method two uses glacial acetic acid. Belinda is interested in if the yield is better when using citric acid, since this method produces more heat and requires more attention. She's done 10 experiments using each method and the yield (calculated as a fraction of the amount of hexamethylene diamine used) rounded off can be seen in the table below. We assume that the measurements are normally distributed and that different batches are independent. We also assume that the variance is the same for both methods.

	Yield										$\overline{x}$	s
Citric acid	55	36	55	64	53	58	55	45	51	40	51.2	8.5088
Acetic acid	50	38	39	40	27	54	47	40	53	35	42.3	8.5641

(a) Perform a test using at least one confidence interval to see if the method using citric acid produces a better yield with confidence level 90%. (2p)

(b) Perform a test to check if it was reasonable to assume that the variances were equal (with significance level 5%). (1p)

4. In MATLAB there is a command `randn(n)` for creating square matrices with random elements from a normal distribution. When testing to generate a growing sequence of matrices and timing the operation for each matrix (for example by using the command `cputime`), the following execution times were obtained.

x	1.0	2.0	3.0	4.0	5.0	6.0	7.0	8.0	9.0	10.0	11.0	12.0	13.0	14.0
y	0.03	0.09	0.21	0.37	0.58	0.84	1.15	1.49	1.90	2.32	2.83	3.36	3.93	4.57

The number x is the number of rows times divided by one thousand (so $x = 2$ means 2000 rows) and y is the execution time (the time it took to generate the matrix in question). Since the number of elements in the matrix grows quadraticly, the following model seems reasonable

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where $\epsilon \sim N(0, \sigma)$ and different measurements are assumed to be independent.

(a) At the 1% level, can you reject the hypothesis that $\beta_1 = 0$? Interpretation? (1p)

(b) Find a prediction interval with degree of confidence 99% for the execution time if $x = 11.5$. (2p)

(c) Estimate the execution time if $x = 20$ using a reasonable estimate based on the model. Is there a problem with using the model for "large" x? (1p)

A helpful mathematician has done the following calculations for you.

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	$-0.2198 \cdot 10^{-3}$	$6.7413 \cdot 10^{-3}$
1	$0.5220 \cdot 10^{-3}$	$2.0676 \cdot 10^{-3}$
2	$23.2692 \cdot 10^{-3}$	$0.1341 \cdot 10^{-3}$

Analysis of variance

	Degrees of freedom	Square sum
REGR	2	29.3755
RES	11	$5.7582 \cdot 10^{-4}$
TOT	13	29.3761

$$(X^T X)^{-1} = \begin{pmatrix} 868.1319 & -239.0110 & 13.7363 \\ -239.0110 & 81.6621 & -5.1511 \\ 13.7363 & -5.1511 & 0.3434 \end{pmatrix} \cdot 10^{-3}.$$

5. Conny is ordering a bunch of seeds from the Netherlands, planning to do some "farming." The guy selling them claims that out of 10 seeds, at least 8 will grow pretty much no matter how much abuse you throw at them. Conny believes this and orders 15 seeds and plants them. After a while, he finds that only 10 has grown. Conny – who considers himself a decent enough statistician – forms the hypothesis test $H_0 : p = 0.8$ versus $H_1 : p < 0.8$ and assumes that the seeding of the seeds is independent.

   (a) Carry out the test at the significance level 5%. (1p)

   (b) What is the power of the test at $p = 0.65$? (1p)

   (c) How many seeds would Conny need to order to obtain a test with a power of 90% at $p = 0.65$ (using the same level of significane)? (2p)

6. Suppose that $\boldsymbol{Y} = (Y_1\ Y_2\ \cdots\ Y_k)^T$, where the components Y_i are normally distributed and independent with the same variance. If $A, B \in \mathbf{R}^{k\times k}$ are constant symmetric matrices such that $A^2 = A$ and $B^2 = B$, prove that $\boldsymbol{Y}^T A\boldsymbol{Y}$ and $\boldsymbol{Y}^T B\boldsymbol{Y}$ are independent if $AB = 0$. (2p)

Solutions

TAMS24/TEN1 2018-10-22

1. Let $x_1, x_2, \ldots, x_n$ be a sample of size n from the Gamma distribution given in the exercise.

   (a) The likelihoodfunction $L(\theta)$ is given by

   $$L(\theta) = \prod_{i=1}^{n} \frac{\theta(\theta x_i)^{v-1} e^{-\theta x_i}}{\Gamma(v)}.$$

   The parameter space is $\Omega_\theta = (0, \infty)$. We were not given any restrictions on θ, but without this assumption it is not clear that we end up with a density function. We form the loglikelihood and take the derivative

   $$\log L(\theta) = \sum_{i=1}^{n} \big(v \log \theta + (v-1) \log x_i - \theta x_i - \log \Gamma(v) \big),$$
   $$\frac{d \log L(\theta)}{d\theta} = \frac{n}{\theta} + \frac{n(v-1)}{\theta} - \sum_{i=1}^{n} x_i.$$

   We're seeking an extremum, so we're looking for points where the derivative is zero:

   $$\frac{n + n(v-1)}{\theta} - n\bar{x} = 0 \qquad \Leftrightarrow \qquad \theta = \frac{v}{\bar{x}},$$

   where $\bar{x}$ is the mean value of the sample. The sign-change for the derivative at the point $\widehat{\theta} = v/\bar{x}$ is $+0-$, so we're dealing with a maximum. It is also clear that $\widehat{\theta} \in \Omega_\theta$ unless all samples are equal to zero, which would be a ridiculous sample.

   (b) We replace $\bar{x}$ by the stochastic quantity $\bar{X}$, where x_j are observations of X_j that are Gamma distributed. We obtain that

   $$E\left(\frac{1}{\widehat{\Theta}}\right) = E\left(\frac{\bar{X}}{v}\right) = \frac{1}{nv} \sum_{i=1}^{n} E(X_i) = \frac{1}{nv} \frac{nv}{\theta} = \frac{1}{\theta},$$

   where we used the hint that told us that the expectation of a Gamma distributed random variable is v/θ.

   Answer: a) The ML-estimate is given by $\widehat{\theta} = v/\bar{x}$. b) See above.

2. Let H_0 be the hypothesis that the data is from a Po(1) variable X and H_1 that this is not true. In total, we have $n = 100$ observations. Suppose that H_0 is true. Then

   $$P(X = j) = \frac{1^j}{j!} e^{-1} = \frac{e^{-1}}{j!},$$

   so we can calculate the last two lines in the following table

$X = ?$	0	1	2	3	4	≥ 5
x_j (frequency)	38	33	26	2	1	0
p_j	0.368	0.368	0.184	0.061	0.015	0.004
np_j	36.8	36.8	18.4	6.1	1.5	0.4

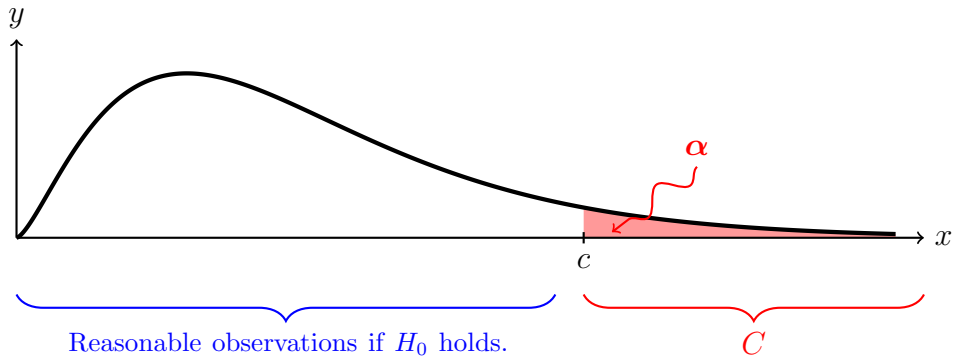
We used that $P(X \geq 5) = 1 - \sum_{j=0}^{4} p_j$ to calculate p_5. We realize here that we have a problem with the last categories since $np_j < 5$. We have to merge these to use a χ^2-test, so let us consider the following categories.

$X = ?$	0	1	2	≥ 3
x_j (frequency)	38	33	26	3
p_j	0.368	0.368	0.184	0.0805
np_j	36.8	36.8	18.4	8.05

The normal test quantity is found in

$$q = \sum_{j=0}^{3} \frac{(x_j - np_j)^2}{np_j} = \frac{(38 - 36.8)^2}{36.8} + \frac{(33 - 36.8)^2}{36.8} + \frac{(26 - 18.4)^2}{18.4} + \frac{(3 - 8.05)^2}{8.05} = 6.7387.$$

If H_0 is true, then q is an observation of $Q \overset{\text{appr.}}{\sim} \chi^2(4-1) = \chi^2(3)$. We reject H_0 if q is large, so we need a critical region C. From a table we find that $c = \chi^2_{0.01}(3) = 11.34$ and we define $C = [c, \infty)$. If $q \geq c$, we reject H_0.



Since $q \notin C$, the conclusion is that we can't reject H_0. So Sture is wrong in rejecting H_0 at this level.

Answer: Sture is wrong. The hypothesis can not be rejected at this level.

3. So the model is that for using citric acid we assume that $X_i \sim N(\mu_1, \sigma^2)$ and for acetic acid that $Y_i \sim N(\mu_2, \sigma^2)$ (same variance). All variables are assumed to be independent.

   (a) We weight together the variances according to the pooled variance:

   $$s^2 = \frac{9s_1^2 + 9s_2^2}{18} = \frac{1}{2}\left(s_1^2 + s_2^2\right).$$

   It now follows that (by Cochran's and Gosset's theorems)

   $$T = \frac{\overline{X} - \overline{Y} - (\mu_1 - \mu_2)}{S\sqrt{\frac{1}{10} + \frac{1}{10}}} \sim t(18),$$

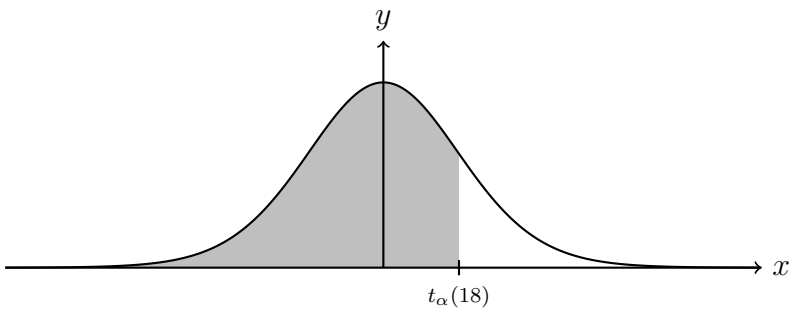
   and

   $$P(T < t_\alpha(18)) = 1 - \alpha,$$

where we can solve the inequality for

$$\overline{X} - \overline{Y} - t_\alpha(18) \cdot \frac{S}{\sqrt{5}} < \mu_1 - \mu_2.$$

We use a one-sided interval since we only want to investigate if $\mu_1 > \mu_2$. From a table, we find that $t_{0.10}(18) = 1.3304$.



As an observation of S, we use $\sqrt{s^2}$, so

$$t_{0.10}(18)\frac{s}{\sqrt{5}} = 1.3304 \cdot 3.8176 = 5.0790.$$

Since $\overline{x} - \overline{y} = 8.9$, the interval is given by

$$\begin{aligned} I_{\mu_1-\mu_2} &= (8.9 - 5.0790,\, \infty) \\ &= (3.821,\, \infty). \end{aligned}$$

We see that 0 is not included in the interval, so we can claim that the citric acid produces a better yield at this significance level. (it is reasonable that $\mu_1 > \mu_2$).

(b) Let

$$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma^2$$

versus

$$H_1 : \sigma_1^2 \neq \sigma_2^2.$$

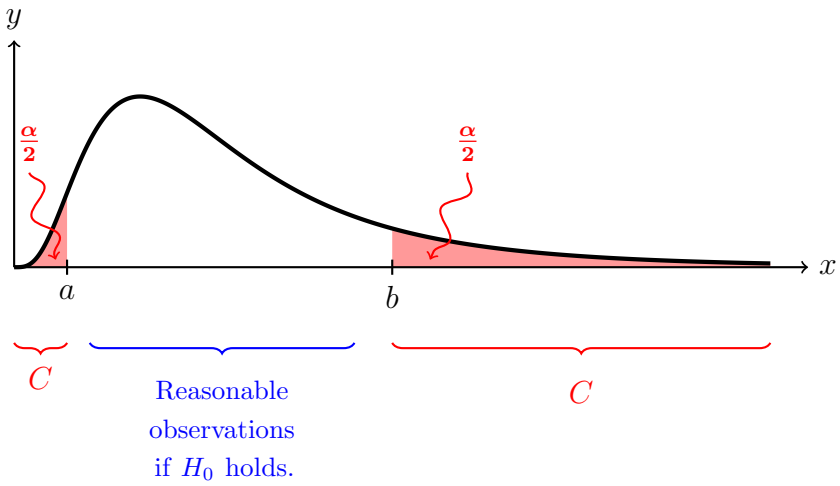
If H_0 is true, then $9S_1^2/\sigma^2 \sim \chi^2(9)$ and $9S_2^2/\sigma^2 \sim \chi^2(9)$. Since these quantities are independent, we have

$$V = \frac{\frac{9S_1^2}{\sigma^2}/(9)}{\frac{9S_2^2}{\sigma^2}/9} = \frac{S_1^2}{S_2^2} \sim F(9,9).$$

We're looking for a critical region C such that

$$\alpha = P(V \in C \,|\, H_0)$$

and we reject H_0 if we obtain an observation in C.



We find a and b from a table such that

$$P(V < a) = P(V > b) = \frac{\alpha}{2} = 0.025,$$

so $b = 4.0260$ and $a = 0.2484$. Since

$$v = \frac{8.5088^2}{8.5641^2} = 0.9871 \notin C$$

we can't reject H_0. The variances might be the same (it is not unreasonable).

Answer: (a) Citric acid produces a better yield. (b) Inconclusive. It is not unreasonable that the variances are the same.

4. We can formulate the problem as a matrix equation $Y = X\beta + \epsilon$, where X is the design matrix

$$X^T = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14 \\ 1 & 4 & 9 & 16 & 25 & 36 & 49 & 64 & 81 & 100 & 121 & 144 & 169 & 196 \end{pmatrix}.$$

The LSE $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ can be obtained from the well-known formula

$$(X^T X)^{-1} X^T y = \begin{pmatrix} -0.2198 \\ 0.5220 \\ 23.2692 \end{pmatrix} \cdot 10^{-3}.$$

We recognize this from the data given in the exercise (and thus it's not a step necessary for the solution). The estimated regression line can be written

$$\widehat{\mu}(x) = (-0.21978 + 0.521978x + 23.269x^2) \cdot 10^{-3}$$

and the estimate value at the point $x = 20$ is obtained as $\widehat{\mu}(20) = \widehat{\boldsymbol{\beta}}^T \cdot (1,\ 20,\ 20^2)^T \approx 9.32$, so that's the answer for the last part of this exercise. But we'll get back to that.

(a) To test if $\beta_1 = 0$, let $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$. Assume that H_0 holds. Then

$$T = \frac{\widehat{\beta}_1 - 0}{S\sqrt{h_{11}}} \sim t(11),$$

where the distribution is clear since H_0 holds. We need a critical region C such that $P(T \in C \,|\, H_0) = 0.05$ and since H_1 is double sided, we choose symmetrically.

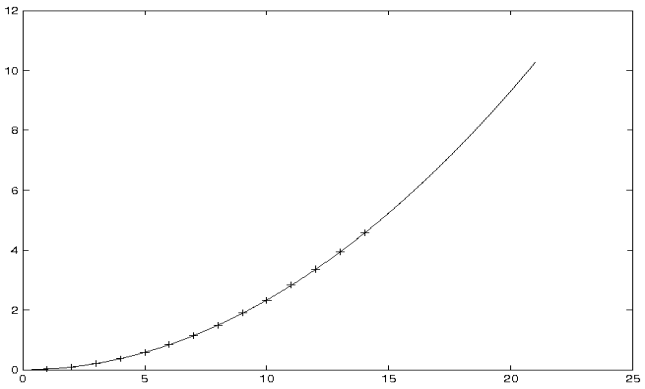
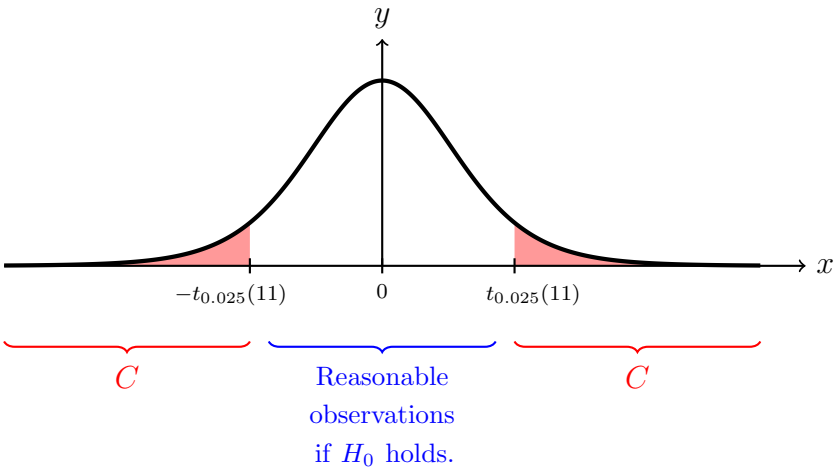


Figure 1: The given measurements match the regression line. If the model holds when $x \gg 14$ is not clear.



We find $t_{\alpha/2}(11) = t_{0.005}(11) = 3.1058$ in a table. An observation of $S\sqrt{h_{11}}$ is given by the standard error $d(\widehat{\beta}_1)$ and thus we find that the observation

$$t = \frac{0.5220 \cdot 10^{-3}}{2.0676 \cdot 10^{-3}} = 0.2525$$

does *not* belong to the critical region. So we can not reject H_0. The coefficient β_1 might very well be zero.

(b) When it comes to finding a prediction interval at $x = 11.5$, we let $u = (1\ 11.5\ 11.5^2)^T$ and Y_0 be an independent random observation at $x = 11.5$. Let $\widehat{\mu_0}$ be the estimate for μ at $x = 11.5$. A well known test quantity is

$$T = \frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(11).$$

We can box in this variable and solve for Y_0:

$$\begin{aligned} -t < T < t \quad &\Leftrightarrow \quad -t < \frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} < t \\ &\Leftrightarrow \quad \widehat{\mu_0} - tS\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}} < Y_0 < \widehat{\mu_0} + tS\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}, \end{aligned}$$

where $t = t_{\alpha/2}(11) = t_{0.005}(11) = 3.1058$. We can now calculate that

$$\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u} = 0.1417,$$

so $\sqrt{1 + \boldsymbol{u}^T (X^T X)^{-1} \boldsymbol{u}} = 1.0685$. As an observation of S, we use

$$s = \sqrt{\frac{5.7582 \cdot 10^{-4}}{11}} = 0.0072.$$

For $\widehat{\mu}_0$, we use the observation $\boldsymbol{u}^T \widehat{\boldsymbol{\beta}} = 3.0831$. Thus we obtain the prediction interval

$$I_{Y_0} = \Big(3.0831 \mp 3.1058 \cdot 0.0072 \cdot 1.0685 \Big) = (3.0592,\ 3.1070).$$

Answer:

(a) No, β_1 could be zero. A possible interpretation is that the first degree term is drowned out in the second algorithmically, meaning that operations are always performed on all elements squarely.

(b) (3.059, 3.107).

(c) The estimated value is 9.32. It is always uncertain to predict values outside the domain from which we've measured. In this case, there would be a gigantic shift in time consumption if the physical RAM memory would run out and the computer moves on to storing things on a hard drive instead. It might stile scale quadratically, but the constants will change. Moreover, if the hard drive runs out of space? What happens then... etc. Don't use a method outside the interval for which we have observations.

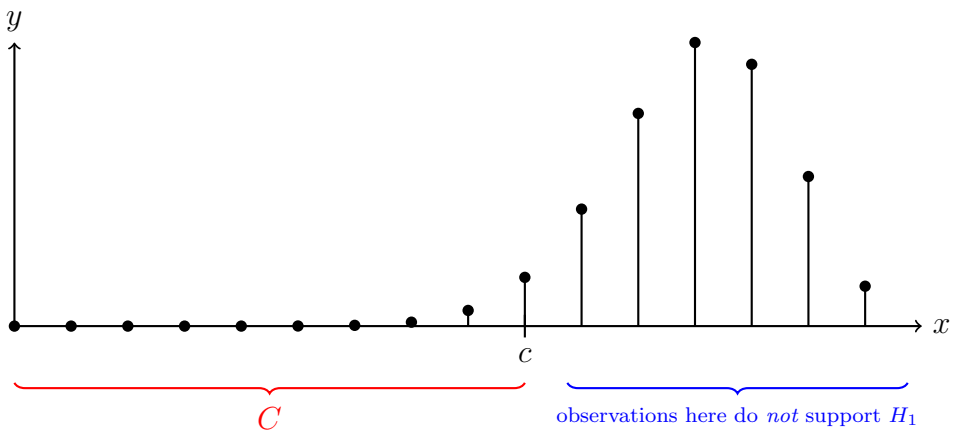
5. (a) We help Conny by performing his hypothesis test. Let X be the number of seeds that grow when planting 15. Then $X \sim \text{Bin}(n, p)$, where $n = 15$ och p is the unknown probability of a seed to grow. We want to test

$$H_0 : p = 0.8$$

versus

$$H_1 : p < 0.8.$$

Given that H_0 is true, we expect the frequency $15 \cdot 0.8 = 12$ for the number of seeds that grow. Is $x = 10$ significantly less? We need the critical region C.



Since

$$p(x) = \binom{15}{x} 0.8^x \cdot 0.2^{15-x},$$

we can calculate that

$$\sum_{x=0}^{8} p(x) = 0.0181 \quad \text{och} \quad \sum_{x=0}^{9} p(x) = 0.0611$$

so it is clear that $c = 8$ is necessary for obtaining a significance level not higher than 5% . Thus,

$$C = \{x \in \mathbf{Z} : 0 \leq x \leq 8\}$$

and our observation $x = 10 \notin C$. Hence we can't reject H_0. The seller might be speaking the truth.

(b) The power at $p = 0.65$ can be calculated straight from the definition:

$$\begin{aligned} h(0.65) &= P(H_0 \text{ rejected} \,|\, p = 0.65) = P(X \in C \,|\, p = 0.65) \\ &= \sum_{x=0}^{8} \binom{15}{x} 0.65^x \cdot 0.35^{15-x} = 0.2452. \end{aligned}$$

(c) We assume that we can approximate X by a normal distribution so that

$$X \overset{\text{appr.}}{\sim} \mathrm{N}(np, np(1-p)).$$

This will be okay if $np(1-p) \geq 10$. We'll need to check that this holds when we're done. Let $c = \Phi^{-1}(0.05)$ and $d = \Phi^{-1}(0.90)$. Then

$$0.05 = \Phi(c) = P(H_0 \text{ rejected} \,|\, p = 0.8) = P(X \leq a \,|\, p = 0.8) \approx \Phi\left(\frac{a - 0.8n}{\sqrt{n \cdot 0.8 \cdot 0.2}}\right)$$

$$\Leftrightarrow \quad c = \frac{a - 0.8n}{\sqrt{n \cdot 0.8 \cdot 0.2}} \quad \Leftrightarrow \quad a = c\sqrt{n \cdot 0.16} + 0.8n$$

and

$$0.9 = P(H_0 \text{ rejected} \,|\, p = 0.65) = P(X \leq a \,|\, p = 0.65) \approx \Phi\left(\frac{a - 0.65n}{\sqrt{n \cdot 0.65 \cdot 0.35}}\right)$$

$$\Leftrightarrow \quad d = \frac{a - 0.65n}{\sqrt{n \cdot 0.2275}} \quad \Leftrightarrow \quad a = d\sqrt{n \cdot 0.2275} + 0.65n$$

Thus

$$c\sqrt{n \cdot 0.16} + 0.8n = d\sqrt{n \cdot 0.2275} + 0.65n \quad \Leftrightarrow \quad \sqrt{n} = \frac{d\sqrt{0.2275} - c\sqrt{0.16}}{0.15} = 8.4614,$$

so $n = 72$ is sufficient. With this n we can find a according to

$$a = d\sqrt{n \cdot 0.2275} + 0.65n = 51.99$$

or

$$a = c\sqrt{n \cdot 0.16} + 0.8n = 52.02.$$

We choose $a = 51$ to be sure that the critical region doesn't become too large. Since $n = 72$ makes $np(1-p) > 10$ for both $p = 0.8$ and $p = 0.65$, our approximation should be okay.

Doing an exact verification, we can see that if $X \sim Bin(72, p)$ *we obtain that*

$$P(X \leq 51 \,|\, p = 0.8) = 0.0406 \quad \textit{and} \quad P(X \leq 51 \,|\, p = 0.65) = 0.8783.$$

Alternate interpretation. One could also interprete the question as using the exactly same significance level we ended up with earlier, i.e., $\alpha = 0.0181$. In this case, letting $c = \Phi^{-1}(0.0181) = -2.0947$ we will obtain that $\sqrt{n} = 9.6609$, so $n = 94$ would be chosen. This leads to $a = 67$.

Answer:

(a) We can't reject H_0.

(b) The power is 0.2452.

(c) $n = 72$ (giving $c = 51$).

6. It is clear that $D = \text{cov}(\boldsymbol{Y})$ is a diagonal matrix since the components are independent. Moreover, since $A^2 = A^T = A$, we can see that

$$\boldsymbol{Y}^T A\boldsymbol{Y} = \boldsymbol{Y}^T A^T A\boldsymbol{Y} = (A\boldsymbol{Y})^T (A\boldsymbol{Y}),$$

and similarly that $\boldsymbol{Y}^T B\boldsymbol{Y} = (B\boldsymbol{Y})^T (B\boldsymbol{Y})$. Let us show that $A\boldsymbol{Y}$ and $B\boldsymbol{Y}$ are independent. The result will then follow since $\boldsymbol{Y}^T A\boldsymbol{Y}$ and $\boldsymbol{Y}^T B\boldsymbol{Y}$ clearly are functions of $A\boldsymbol{Y}$ and $B\boldsymbol{Y}$, respectively. Since $\boldsymbol{Y}$ is normally distributed, this is also true for $A\boldsymbol{Y}$ and $B\boldsymbol{Y}$. Thus it is enough to show that the variables are uncorrelated to obtain independence. The covariance is given by

$$\begin{aligned}
\text{cov}(A\boldsymbol{Y}, B\boldsymbol{Y}) &= E(A\boldsymbol{Y}(B\boldsymbol{Y})^T) - E(A\boldsymbol{Y})E(B\boldsymbol{Y})^T \\
&= AE(\boldsymbol{Y}\boldsymbol{Y}^T)B^T - AE(\boldsymbol{Y})(BE(\boldsymbol{Y}))^T \\
&= A\left(E(\boldsymbol{Y}\boldsymbol{Y}^T) - E(\boldsymbol{Y})E(\boldsymbol{Y})^T\right) B^T = A\left(\text{cov}(\boldsymbol{Y})\right) B^T = ADB = DAB = 0,
\end{aligned}$$

if $D = \sigma^2 I$.

Answer: See above.

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk statistik

Kurskod: TAMS24
Provkod: TEN1
Datum: 2019-01-09
Institution: MAI

Exam in Statistics

TAMS24/TEN1 2019-01-09

You are permitted to bring:

- a calculator (no computer);
- Formel- och tabellsamling i matematisk statistik (from MAI);
- Formel- och tabellsamling i matematisk statistik, TAMS65;
- TAMS24: Notations and Formulas (by Xiangfeng Yang).

Grading (sufficient limits): 8-11 points giving grade 3; 11.5-14.5 points giving grade 4; 15-18 points giving grade 5. Your solutions need to be complete, well motivated, carefully written and concluded by a clear answer. Be careful to show what is random and what is not. Assumptions you make need to be explicit. The exercises are in number order.

Solutions can be found on the homepage a couple of hours after the finished exam.

1. A reasonable question is which one of Morbid Angel's first bunch of studio albums is the best (so no live albums and to avoid confusion not the Abominations of Desolation album[1]). The following data was collected from two sources on the internet.

Album Title	Web page	
	Nuclear War Now!	Metalstorm.net
Altars of Madness	67	82
Blessed Are The Sick	18	34
Covenant	11	32
Domination	1	21
Formulas Fatal To The Flesh	6	4
Gateways to Annihilation	1	10
Heretic	0	2

   Use a suitable test with significance level 0.01 to see if there is a difference between opinions on the two sites. (2p)

2. Lina is experimenting with water cooling for her computers. She's measured temperatures of the cpu:s in 5 different computers (those who survived the experiment), first with conventional cooling and then after switching to water cooling.

[1] Since it's obviously the best.

	Temperature				
Computer:	C-1	C-2	C-3	C-4	C-5
Conventional cooling	55	36	55	64	53
Water cooling	50	38	39	50	44

We assume that the temperatures are normally distributed and that different computers are independent.

(a) Find 95% confidence intervals for the expected temperatures with conventional cooling and water cooling, respectively. (2p)

(b) Perform a test for if there is a difference in the expected temperature between the cooling techniques at the level 5%. (2p)

(c) Find a confidence interval $I_{\sigma^2} = [0, a)$ for the variance of the temperature using water cooling. Use the degree of confidence 90%. (1p)

3. In statistics, one frequently works with *stochastic processes*. One such example could be expressed as $X(n)$ for $n = 1, 2, 3, \ldots$, where $X(n)$ is a random variable for each n. Let one such process $X(n)$ satisfy the following. It has expectation 0 for every n (that is, $E(X(n)) = 0$) and if $Y(n)$ is the random vector $Y(n) = (\, X(n),\ X(n-1),\ X(n-2)\,)^T$, then the covariance matrix is given by

$$C_{Y(n)} = E(Y(n)Y(n)^T) = \begin{pmatrix} 2 & 1 & 0 \\ 1 & 2 & 1 \\ 0 & 1 & 2 \end{pmatrix}$$

for every $n = 3, 4, 5, \ldots$.

Find a linear predictor $\widehat{X}(n) = aX(n-1) + bX(n-2)$ of $X(n)$ that minimizes the quadratic error. In other words, find a and b such that $E((\widehat{X}(n) - X(n))^2)$ is minimal. (2p)

4. Crawford Tillinghast has built a machine that enables people to see and interact with alternate dimensions. It works by stimulating the pineal gland in the brain by means of resonance waves. While building his machine, he measured the frequency of the waves as a function of the voltage he applied, and he also took note of if the measurement was made at night (represented by 0) or during the day (represented by 1).

Frequency (f)	1.86	2.41	3.26	3.88	4.64	5.50	6.40
Voltage (v)	1	2	3	4	5	6	7
Day/Night (u)	0	0	1	0	0	1	1

Crawford believes that the frequency depends linearly on the voltage, but he isn't sure that the time of day is important (but he gets his most spectacular results at night). He considers the following two models:

$$\text{Model 1:} \quad F = \beta_0 + \beta_1 v + \epsilon$$

and

$$\text{Model 2:} \quad F = \beta_0 + \beta_1 v + \beta_2 u + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and different measurements are assumed to be independent. The following calculations has already been carried out.

Model 1:

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	0.9671	0.0984
1	0.7564	0.0220

Analysis of variance

	Degrees of freedom	Square sum
REGR	1	16.0212
RES	5	0.0678
TOT	6	16.0889

Model 2:

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	0.9866	0.0923
1	0.7370	0.0250
2	0.1363	0.1009

Analysis of variance

	Degrees of freedom	Square sum
REGR	2	16.0424
RES	4	0.0466
TOT	6	16.0889

(a) Is the term in model 2 corresponding to day/night meaningful? Carry out a test at the 1%-level. What is the interpretation of your result? (2p)

(b) Find a 95% confidence interval for β_1 using model 2. (1p)

5. In a game of *Death Adder Roulette*, played out in the Australian outback, people take turns in trying to pet a venomous snake (traditionally a death adder) on the head. The game is played until someone is bitten. Assume that the probability of being bitten is constant.

(a) Assume that a person is bitten at try number n. Find a reasonable (using n) point estimate for p and calculate the expectation for the estimator. Is the estimator unbiased? (2p)

(b) One participant claims that the current snake i feisty so that $p = 0.4$. In one game, the first person was bitten at the fifth try. Test the hypothesis $H_0 : p = 0.4$ versus $H_1 : p < 0.4$ at the significance level 5%. (1p)

(c) What is the power of the test at $p = 0.2$? (1p)

6. Suppose that $\boldsymbol{Y} = (Y_1\ Y_2\ \cdots\ Y_k)^T \sim N(\mathbf{0}, I_k)$, where I_k is the $k \times k$ identity matrix. If $A \in \mathbf{R}^{k\times k}$ is such that A is symmetric, the column rank of A is l $(0 < l \leq k)$, and $A^2 = A$, derive the distribution for $\boldsymbol{Y}^T A \boldsymbol{Y}$. (2p)

Solutions

TAMS24/TEN1 2019-01-09

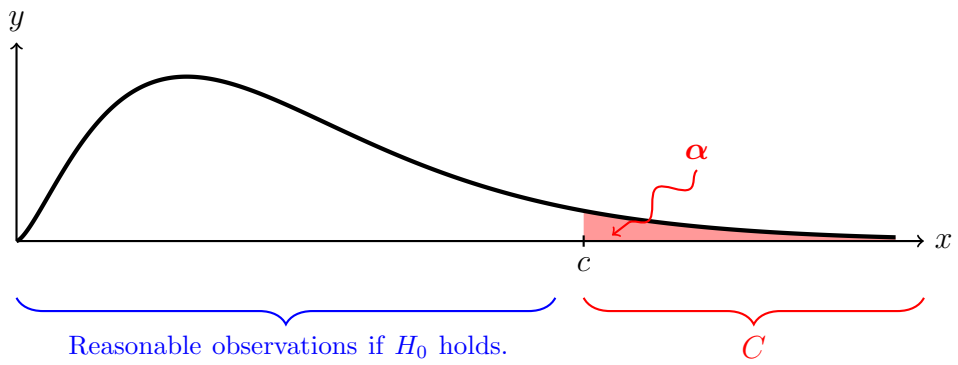
1. Let H_0 be the hypothesis that the data is homogeneous between the two sites and H_1 that this is not true. In total, we have $n = 289$ observations. We can directly see that the last four albums will have too small $n_i\widehat{p}_j$ (significantly less than 5), so we have to combine these to obtain a usable test. Note that this changes what we actually test, but it's the best we can do using the tools from this course. We can calculate the following from the data given.

Album Title	Web page		Sum	$\widehat{p}_j$
	Nuclear War Now!	Metalstorm.net		
Altars of Madness	67	82	149	0.516
Blessed Are The Sick	18	34	52	0.180
Covenant	11	32	43	0.149
D–H	8	37	45	0.156
n_i	104	185	289	

The usual test quantity is found in

$$\begin{aligned} q &= \sum_{i=0}^{1}\sum_{j=0}^{3}\frac{(N_{ij}-n_i\widehat{p}_j)^2}{n_i\widehat{p}_j} = \frac{(67-53.62)^2}{53.62} + \frac{(82-95.38)^2}{95.38} + \frac{(18-18.72)^2}{18.72} \\ &\quad + \frac{(34-33.29)^2}{33.29} + \frac{(11-15.47)^2}{15.47} + \frac{(32-27.53)^2}{27.53} \\ &\quad + \frac{(8-16.19)^2}{16.19} + \frac{(37-28.81)^2}{28.81} \\ &= 13.76. \end{aligned}$$

If H_0 is true, then q is an observation of $Q \overset{\text{appr.}}{\sim} \chi^2((2-1)(4-1)) = \chi^2(3)$. We reject H_0 if q is large, so we need a critical region C of the form $C = [c, \infty)$. From a table we find that $c = \chi^2_{0.01}(3) = 11.34$. If $q \geq c$, we reject H_0.



Since $q \in C$, the conclusion is that we reject H_0. There is very likely a difference in opinions between the two sites.

Answer: There is a difference.

2. (a) Let X_i be the temperatures with conventional cooling and Y_i the temperatures with water cooling. Assume that $X_i \sim N(\mu_X, \sigma_X^2)$ and that $Y_i \sim N(\mu_Y, \sigma_Y^2)$, where μ_X and μ_Y are the expected temperatures using the different cooling techniques. We can not assume that the variance is the same or that X_i and Y_i are independent, but different X_i and different Y_i are independent. We do not know that this model is true (there might be different expected temperatures for the different computers), but it's the best we can do to answer the question. Another interpretation is that it is the mean temperatures we're interested in.

It now follows that (by Cochran's and Gosset's theorems)

$$T_X = \frac{\overline{X} - \mu_X}{S/\sqrt{5}} \sim t(4),$$

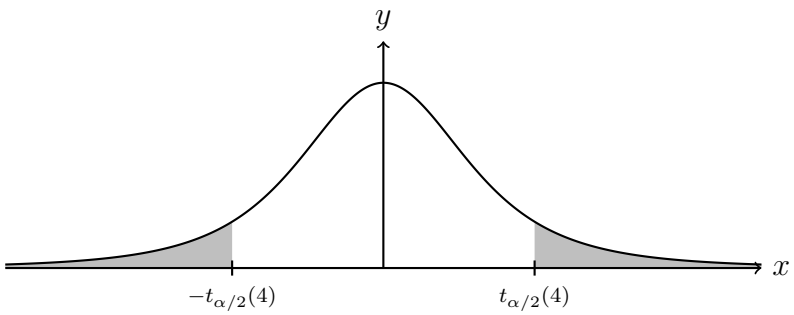
and

$$P(-t_{\alpha/2}(4) < T_X < t_{\alpha/2}(4)) = 1 - \alpha,$$

where we can solve the inequality for

$$\overline{X} - t_{\alpha/2}(4) \cdot \frac{S}{\sqrt{5}} < \mu_X < \overline{X} + t_{\alpha/2}(4) \cdot \frac{S}{\sqrt{5}}.$$

From a table, we find that $t_{0.025}(4) = 2.7764$.



As an observation of S_X, we use $\sqrt{s_X^2}$, so

$$t_{0.025}(4)\frac{s}{\sqrt{5}} = 2.7764 \cdot \frac{10.2127}{2.2361} = 12.6808.$$

Since $\overline{x} = 52.6$, the interval is given by

$$I_{\mu_X} = (39.9,\ 65.3).$$

Analogously, we find a confidence interval for μ_Y in

$$I_{\mu_Y} = (37.1,\ 51.4).$$

(b) To obtain a significant result, we can not use the intervals derived in (a) for several reasons. First, the intervals are not independent (at least we can't be sure). Secondly, the simultaneous degree of confidence will be wrong compared to what we're asked to do in this part.

The model we need to use is samples in pairs.

If x_i is the temperature before introducing water cooling and y_i the temperature after, we assume that x_i are observations of $X_i \sim \mathrm{N}(\mu_i, \sigma_1^2)$ and y_i from $Y_i \sim \mathrm{N}(\mu_i + \Delta, \sigma_2^2)$. Define $Z_i = Y_i - X_i \sim \mathrm{N}(\Delta, \sigma^2)$. We consider the sequence $z_i = y_i - x_i$ as observations of Z_i. Note that the variables Z_i are independent since we assumed that different computers are independent.

	Temperature difference				
z_i	5	-2	16	14	9

We can now calculate $s = 7.2319$ and $\overline{z} = 8.4$. Moreover, $n - 1 = 4$ and $\alpha = 0.05$, so $t_{\alpha/2}(4) = t_{0.025}(4) = 2.7764$. Thus,

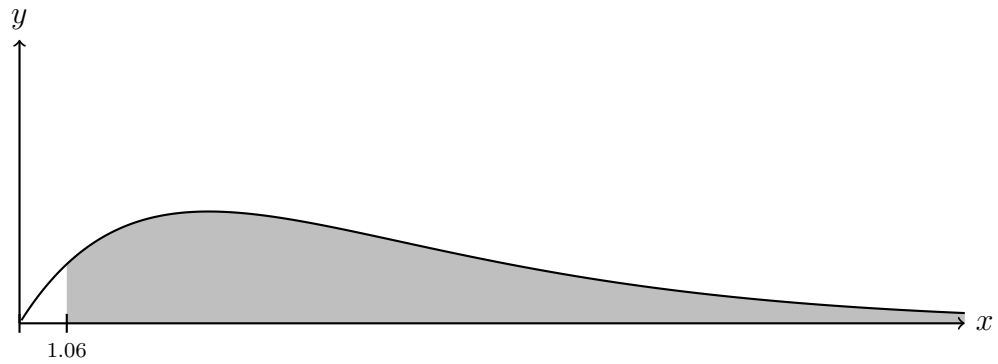
$$I_\Delta = (8.4 - 2.7764 \cdot 7.2319/\sqrt{5},\ 8.4 + 2.7764 \cdot 7.2319/\sqrt{5}) = (-0.58,\ 17.4).$$

Since $0 \in I_\Delta$, we can't reject the hypothesis that $\Delta = 0$. It is not clear that there is a difference.

(c) This is a similar situation to (a), where we have to assume that the temperatures are from the same distribution $N(\mu_Y, \sigma^2)$ (or consider the mean temperature). We define

$$V = \frac{4\,S^2}{\sigma^2} \sim \chi^2(4).$$

From a table we find c such that $P(c < V) = 0.90$ by choosing $c = \chi^2_{0.10}(4) = 1.064$.



We solve for σ^2:

$$c < \frac{4\,S^2}{\sigma^2} \quad \Leftrightarrow \quad \sigma^2 < \frac{4\,S^2}{c}$$

and use $s^2 = 33.2$ as the estimate for S^2, leading to the confidence interval

$$I_{\sigma^2} = (0,\ 124.9)\,.$$

Answer:

(a) $I_{\mu_X} = (39.9,\ 65.3)$ and $I_{\mu_Y} = (37.1,\ 51.4)$.

(b) Inconclusive. There might not be a difference.

(c) $I_{\sigma^2} = (0,\ 124.9)$.

3. Let $Z = \widehat{X}(n) - X(n)$. Then $Z = AY(n)$, where $A = (\,-1,\ a,\ b\,)$. Thus,

$$\begin{aligned} E(Z^2) &= V(Z) + E(Z)^2 = AC_{Y(n)}A^T + 0 \\ &= \cdots = 2 - 2a + 2a^2 + 2ab + 2b^2 =: f(a,b). \end{aligned}$$

We seek a and b that minimizes $f(a,b)$. Letting $\nabla f = 0$, we find that

$$\begin{cases} f'_a(a,b) = -2 + 4a + 2b = 0 \\ f'_b(a,b) = 2a + 4b = 0 \end{cases}$$

Solving the system of equations, we obtain $a = 2/3$ and $b = -1/3$. Is this a minimum? Calculating the derivatives of order two, we have $f''_{aa} = f''_{bb} = 4$ and $f''_{ab} = 2$. Looking at the quadratic form,

$$Q(h,k) = 4k^2 + 4hk + 4k^2 = 4(k + h/2)^2 + 3h^2,$$

we see that it is positively definite. Hence this is indeed a minimum.

Answer: The linear predictor is given by

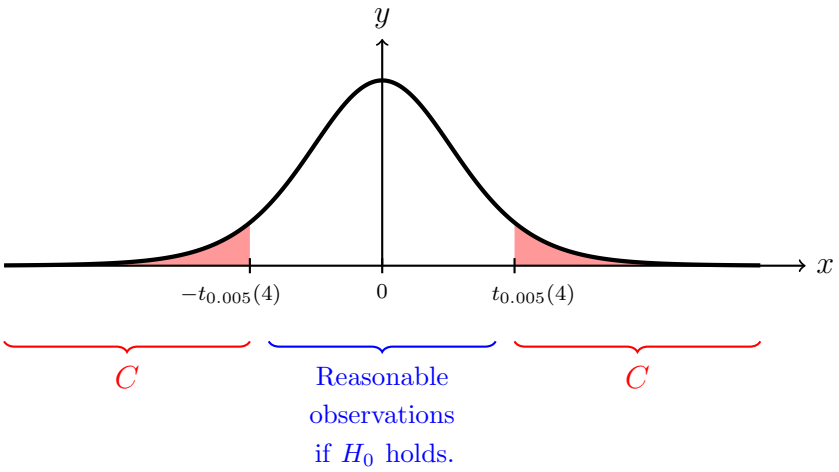
$$\widehat{X}(n) = \frac{2}{3}X(n-1) - \frac{1}{3}X(n-2).$$

4. (a) We can perform this test in several different ways. We can test whether $\beta_2 = 0$ in model 2 directly or we can compare model 1 and model 2 and see if model 2 is significantly better.

Alternative 1. To test if $\beta_2 = 0$, let $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 \neq 0$. Assume that H_0 holds. Then

$$T = \frac{\widehat{\beta}_2 - 0}{S\sqrt{h_{22}}} \sim t(4),$$

where the distribution is clear since H_0 holds. We need a critical region C such that $P(T \in C \,|\, H_0) = 0.01$ and since H_1 is double sided, we choose symmetrically.



We find $t_{\alpha/2}(4) = t_{0.005}(4) = 4.6041$ in a table. An observation of $S\sqrt{h_{22}}$ is given by the standard error $d(\widehat{\beta}_2)$ and thus we find that the observation

$$t = \frac{0.1363}{0.1009} = 1.35$$

does *not* belong to the critical region. So we can not reject H_0. The coefficient β_2 might very well be zero.

Alternative 2.

We have model 1:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

and model 2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon.$$

We can test if the second model is significantly better by testing whether $\beta_2 = 0$ in a slightly different way.

Let

$$H_0 : \beta_2 = 0,$$

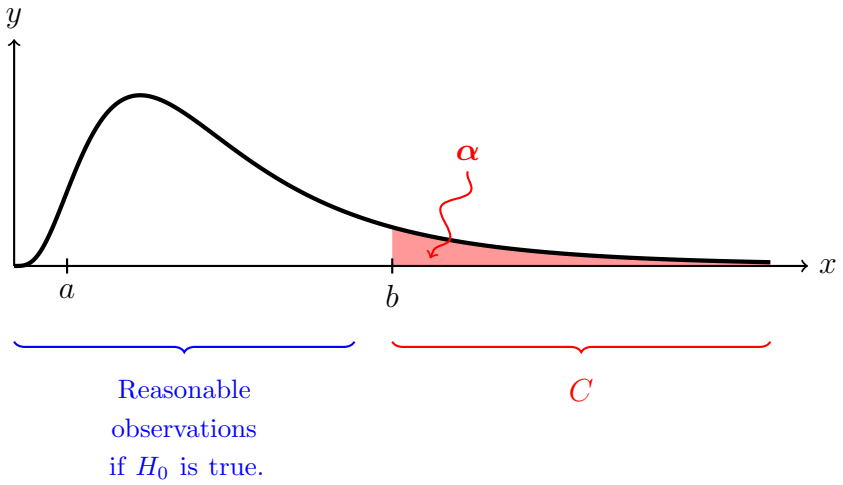
and

$$H_1 : \beta_2 \neq 0.$$

If H_0 is true, then $\boldsymbol{Y} \sim N(X_1\boldsymbol{\beta_1}, \sigma^2 I)$, so

$$W = \frac{(\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)})/1}{\mathrm{SS}_{\mathrm{E}}^{(2)}/4} \sim F(1,4) \quad \text{if } H_0 \text{ is true}$$

since this is a quotient of independent χ^2 variables. If H_0 is not true, then W will tend to grow large. The critical domain is given by $C =]c, \infty[$ for some $c > 0$.



From the table we find that $c = 21.1977$. The observation of W is found as

$$w = \frac{(0.0678 - 0.0466)/1}{0.0466/4} = 1.82,$$

so clearly $w \notin C$. We can not reject the null hypothesis.

(b) We wish to find a confidence interval for β_1 using model 2. We know that

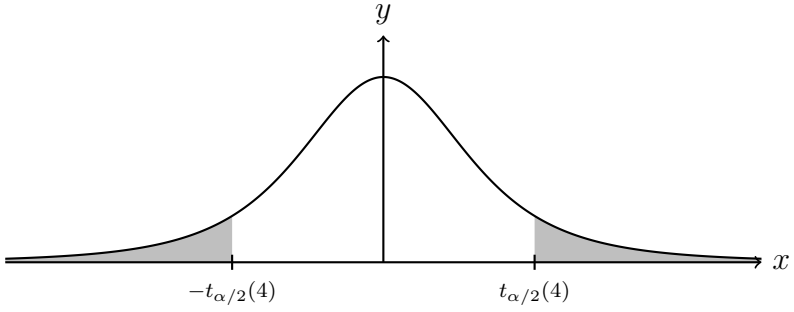
$$T = \frac{\widehat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(4).$$

So

$$P(-t_{\alpha/2}(4) < T < t_{\alpha/2}(4)) = 1 - \alpha,$$

where we can solve the inequality for

$$\widehat{\beta}_1 - t_{\alpha/2}(4) \cdot S\sqrt{h_{11}} < \beta_1 < \widehat{\beta}_1 + t_{\alpha/2}(4) \cdot S\sqrt{h_{11}}.$$



From a table, we find that $t_{0.025}(4) = 2.7764$. An observation of $S\sqrt{h_{11}}$ is given by the standard error $d(\widehat{\beta}_1) = 0.025$ and thus we find the confidence interval

$$I_{\beta_1} = \Big(\widehat{\beta}_1 - 2.7764 \cdot 0.025,\, \widehat{\beta}_1 + 2.7764 \cdot 0.025\Big) = (0.67,\, 0.81).$$

Answer:

(a) A significance test shows that we can't conclude that $\beta_2 \neq 0$ at the significance level 1%. The conclusion is that we really don't know.

(b) $(0.67,\ 0.81)$.

5. (a) A reasonable estimate that is fairly obvious is to let $\widehat{p} = x^{-1}$, where x is the observation of the number of trials it takes for the snake to bite someone. We note that the assumptions lead to the conclusion that $X \sim \mathrm{Ffg}(p)$. If the estimate $\widehat{p} = x^{-1}$ is not obviously reasonable, we can show that this is actually the MLE.

The likelihood-function $L(p)$ is given by

$$L(p) = p(1-p)^{x-1},$$

where x is the observation described above and p is the unknown probability. We only have one probability function to work with, so there's no product of n different probability functions. The parameter space is $\Omega_p = (0, 1)$ (the extreme cases at $p = 0$ and $p = 1$ are not very interesting). We form the log-likelihood and take the derivative with respect to p (remember that x is fixed):

$$\begin{aligned} \log L(p) &= \log p + (x-1)\log(1-p), \\ \frac{d\log L(p)}{dp} &= \frac{1}{p} - \frac{x-1}{1-p}. \end{aligned}$$

We're seeking an extremum, so we're looking for points where the derivative is zero:

$$\frac{1}{p} - \frac{x-1}{1-p} = 0 \qquad \Leftrightarrow \qquad p = \frac{1}{x}.$$

The sign-change for the derivative at the point $\widehat{p} = 1/x$ is $+0-$, so we're dealing with a maximum. It is also clear that $\widehat{p} \in \Omega_p$ since $x \geq 1$.

The expectation of the estimator can be calculated as follows (remember the second course in single variable analysis):

$$E(\widehat{P}) = E(X^{-1}) = \sum_{x=1}^{\infty} x^{-1} p_X(x) = \sum_{x=1}^{\infty} x^{-1} p(1-p)^{x-1} = \frac{p}{1-p} \sum_{x=1}^{\infty} \frac{(1-p)^x}{x}.$$

Let $f(t) = \displaystyle\sum_{k=1}^{\infty} \frac{t^k}{k}$. We can calculate this series by observing that

$$f(t) = \sum_{k=1}^{\infty} \frac{t^k}{k} = \sum_{k=1}^{\infty} \int_0^t u^{k-1}\, du = \int_0^t \left(\sum_{k=1}^{\infty} u^{k-1} \right) du = \int_0^t \frac{1}{1-u}\, du = -\ln(1-t),$$

provided that $0 < t < 1$ (where the series is absolutely convergent). Thus we have shown that

$$E(\widehat{P}) = \frac{p f(1-p)}{1-p} = \frac{-p\ln p}{1-p} \neq p,$$

so the estimator is *not* unbiased.

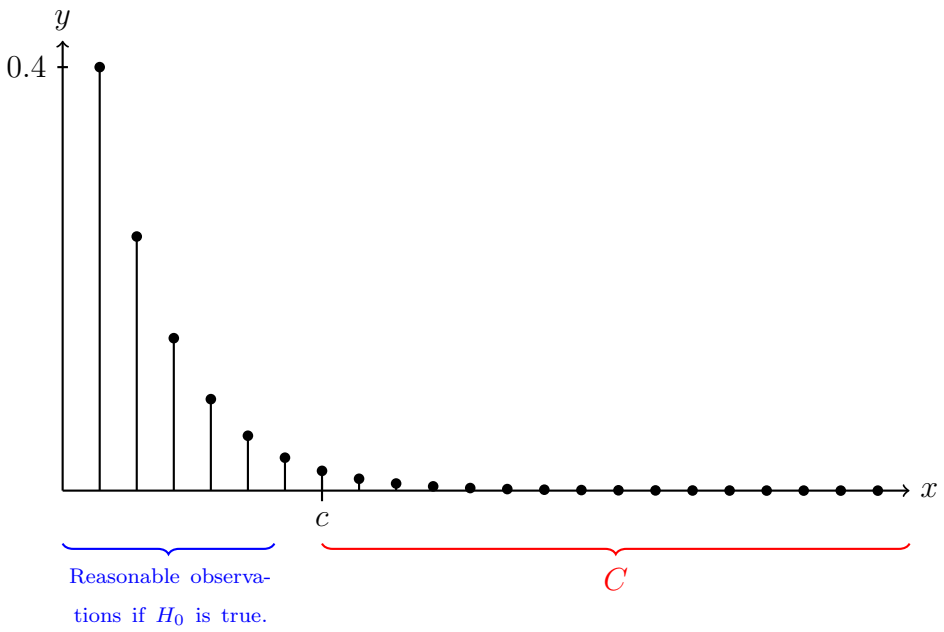
(b) Let X be the number of trials it takes for someone to finally get bitten. We concluded above that $X \sim \text{Ffg}(p)$, where p is the unknown probability of a bite. We want to test

$$H_0 : p = 0.4$$

versus

$$H_1 : p < 0.4.$$

Given that H_0 is true, we expect that it takes $1/0.4 = 2.5$ times to end the game. Is $x = 5$ significantly greater? Large observations indicate that the probability is low. We need the critical region C.



Since

$$p(x) = p(1-p)^{x-1},$$

we can calculate that

$$\begin{aligned} P(X \geq x) &= \sum_{k=x}^{\infty} p(1-p)^{k-1} = p(1-p)^{x-1} \sum_{k=0}^{\infty} (1-p)^k \\ &= p(1-p)^{x-1} \frac{1}{1-(1-p)} = (1-p)^{x-1}. \end{aligned}$$

Testing values for x we find that $P(X \geq 7) \leq 0.05$ but $P(X \geq 6) > 0.05$. So

$$C = \{x \in \mathbf{Z} : x \geq 7\}$$

and our observation $x = 5 \notin C$. Hence we can't reject H_0. The snake might be feisty to a value of $p = 0.4$.

(c) The power at $p = 0.2$ can be calculated straight from the definition:

$$\begin{aligned} h(0.2) &= P(H_0 \text{ rejected} \,|\, p = 0.2) = P(X \in C \,|\, p = 0.2) \\ &= \sum_{x=7}^{\infty} 0.2 \cdot 0.8^{x-1} = 0.262. \end{aligned}$$

Answer: (a) $\widehat{P} = \dfrac{1}{x}$; not unbiased. (b) We can't reject H_0. (c) The power is 0.262.

6. Since A is a symmetric matrix, there exists an orthonormal basis where A is a diagonal matrix. In other words, there is an orthonormal matrix C such that $A = CDC^T$. Let $\boldsymbol{Z} = C^T\boldsymbol{Y}$. Now, since $A^2 = A$, the only possible eigenvalues of A are 0 and 1. These are the values on the diagonal of D. We assume that these are in decreasing order $1, 1, \ldots, 1, 0, \ldots, 0$. The rank of A is l, so there are precisely l ones. Now,

$$\begin{aligned}\boldsymbol{Y}^T A\boldsymbol{Y} &= \boldsymbol{Y}^T CDC^T\boldsymbol{Y} = (C\boldsymbol{Z})^T CDC^T C\boldsymbol{Z} \\ &= \boldsymbol{Z}^T C^T CDC^T C\boldsymbol{Z} = \boldsymbol{Z}^T D\boldsymbol{Z},\end{aligned}$$

since $C^T C = I$. The fact that D is of the form described above shows that

$$\boldsymbol{Z}^T D\boldsymbol{Z} = \sum_{j=1}^{l} Z_j^2.$$

We can also see that the components of $\boldsymbol{Z}$ are independent since

$$\operatorname{cov}(\boldsymbol{Z}) = \operatorname{cov}(C^T\boldsymbol{Y}) = C^T \operatorname{cov}(\boldsymbol{Y})C = C^T C = I$$

due to the fact that $\operatorname{cov}(\boldsymbol{Y}) = I$.

We have thus shown that $\boldsymbol{Y}^T A\boldsymbol{Y}$ can be expressed as a sum of l squares of independent $N(0,1)$-distributed variables. This implies that

$$\boldsymbol{Y}^T A\boldsymbol{Y} \sim \chi^2(l).$$

Answer: $\boldsymbol{Y}^T A\boldsymbol{Y} \sim \chi^2(l)$.

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk statistik

Utbildningskod: TAMS24
Modul: TEN1
Datum: 2019-08-23
Institution: MAI

Exam in Statistics

TAMS24/TEN1 2019-08-23

You are permitted to bring:

- a calculator (no computer);
- Formel- och tabellsamling i matematisk statistik (from MAI);
- Formel- och tabellsamling i matematisk statistik, TAMS65;
- TAMS24: Notations and Formulas (by Xiangfeng Yang).

Grading (sufficient limits): 8-11 points giving grade 3; 11.5-14.5 points giving grade 4; 15-18 points giving grade 5. Your solutions need to be complete, well motivated, carefully written and concluded by a clear answer. Be careful to show what is random and what is not. Assumptions you make need to be explicit. Approximations are allowed if reasonable and clearly motivated. The exercises are in random order.

Solutions can be found on the homepage a couple of hours after the finished exam.

Once upon a time, in a land far far away, Rick and Gary started a flamingo farm that was called Exodus. When starting out, Exodus was filled up by a large amount of flamingos that were obtained from a woman living in the swamps of Louisiana. The farm was built in southern Florida in a suitable habitat for flamingos. Both Gary and Rick quickly became rather proficient in the art of breeding flamingos. Aiming to export flamingos both to the animal parks of the world and to private citizens, Rick and Gary proudly produced commercials exclaiming their competence in flamingo breeding. Before not too long, they got a call from their very first customer. Things did not turn out the way they expected...

1. A Danish doughnut company – with a secret recipe – wants to buy large amounts of flamingos each month for some reason. They are prepared to pay a certain amount per kilogram of flamingos, so heavier flamingos render more profit. Gary picks out a random sample of flamingos and measures their weight:

   2.69 2.90 3.23 3.52 2.65 3.71 3.46 3.05

   Assume that the samples are independent and from a normal distribution with variance 0.0625 and an unknown expectation μ.

   (a) Test the hypothesis $H_0 : \mu = 3.0$ against $H_1 : \mu \neq 3.0$ at the significance level 5%. (2p)

   (b) What is the power of this test at $\mu = 3.1$? (1p)

   (c) What is the highest level of confidence we can choose when using this sample and still reject H_0? Is it reasonable to use this calculation to choose the significance level of the test you want to perform? (2p)

2. When using the results from the previous exercise to decide what to charge the Danish company, things did not turn out exactly as calculated. To avoid a faster disaster, Gary and Rick decided to not assume that the variance is known. Using the same sample as in the previous exercise, answer the following questions.

   (a) Test the assumption that the variance actually is equal to 0.0625 against the alternate hypothesis that the variance is greater. Use the significance level 0.05. (2p)

   (b) Assume that the variance is unknown. Find a confidence interval for μ with 99% degree of confidence. (1p)

3. It turns out that in the contract with the Danish company, there was some fine print detailing that the flamingos were to be slaughtered prior to shipping. Obviously upset, Rick and Gary devised a plan for euthanizing the flamingos as humanly as possible by means of the FLAMINGO DECAPITATOR 2000$^{\text{TM}}$ (a shovel headed killing machine). The blades are very sharp, but need additional sharpening after a certain time to keep the efficiency of the strike of the beast.

   The distributor of the blades (a company called *metal command*) claim that the time until sharpening is necessary (assuming a certain prescribed use) is exponentially distributed with the expectation 1.0 days. Rick and Gary puts this to the test using 50 identical machines in parallel (and in exactly the same way) over the course of 2.5 days. They take note every 6 hours of how many machines that has been taken out for sharpening (these machines are then kept out of circulation to not interfer with the measurements).

Time (hours)	< 6	< 12	< 18	< 24	< 30	< 36	< 42	< 48	< 54	< 60
Frequency:	11	20	26	32	36	39	42	44	46	47

   Use a suitable test with significance level 10% to see if we can reject the hypothesis that the samples are from an $\text{Exp}(\mu = 1.0)$-distribution. (2p)

4. A company called *pleasures of the flesh* got into contact and offered to sell a growth hormone specifically tailored to birds of a similar type as flamingos. The company claimed that the size of the flamingos was linearly dependent on the amount of hormone administered. An experiment to investigate a reasonable dosage was carried out, but when studying residual plots from a linear regression there were hints of something quadratic.

Size (kg)	2.1344	2.3870	2.5861	2.8209	3.0492	3.2521	3.6765	4.0815
Hormone (mg/kg)	0.1250	0.2500	0.3750	0.5000	0.6250	0.7500	0.8750	1.0000

Consider the following two models:

$$\text{Model 1:} \quad Y = \beta_0 + \beta_1 x + \epsilon$$

and

$$\text{Model 2:} \quad Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and different measurements are assumed to be independent. The following calculations has already been carried out.

Model 1:

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	1.8036	0.0784
1	2.1242	0.1241

Analysis of variance

	Degrees of freedom	Square sum
REGR	1	2.9610
RES	6	0.0607
TOT	7	3.0217

Model 2:

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	2.0456	0.0813
1	0.9627	0.3315
2	1.0324	0.2876

Analysis of variance

	Degrees of freedom	Square sum
REGR	2	3.0047
RES	5	0.0170
TOT	7	3.0217

(a) Is the term in model 2 corresponding to x^2 meaningful? Carry out a test at the 5%-level. What is the interpretation of your result? (2p)

(b) Find a 99% confidence interval for β_1 using model 2. Does it differ from the corresponding confidence interval for β_1 using model 1? How do we interpret this? (1p)

5. Exodus runs into a problem of a certain species of fish that competes with the flamingos for food. The first idea for a solution was based on a chemical method called *Chemi-kill*, but due to fears of the effect on the flamingos this plan was scrapped. Fortunately, their close friend Susan stops by and claims that someone told her in a dream that introducing piranhas to the habitat would solve the problem.

   Gary contacts a Peruvian specialist Maria, who claims that there are two particularly ferocious types of red-bellied piranhas. Rick and Gary imports an equal amount of both types and devise an experiment where two identical tanks are filled with the different types of piranhas and 300 exemplars of the problem fish in each tank. What followed was a lesson in violence, where the starving piranhas went to attack. They stop the experiment after a day has passed and takes a count of the remaining problem fish. In the first tank there were 200 left and in the second 180. Let p_1 be the probability that a fish is eaten in the first tank and p_2 be the probability that a fish is eaten in the second tank.

   (a) Propose unbiased point estimators for p_1, p_2, and $p_2 - p_1$. Then find 95% confidence intervals for p_1, p_2 and $p_2 - p_1$. Can we reject that they're equally ferocious at this level? (2p)

   (b) After another call to Maria, she tells them that the second type might be a bit more aggressive. At the significance level 5%, can we reject the hypothesis that the types of piranhas are equally ferocious using the alternate hypothesis that the second type (corresponding to p_2) is more ferocious instead? (1p)

6. At the Exodus farm, a test is planned to verify certain conditions. To understand the test, Rick and Gary are going through the proof but got stuck at the following part. Let $\boldsymbol{p} = (p_1, p_2, \ldots, p_k)^T$ be a probability vector, where $k \geq 2$ is an integer. Suppose that $\boldsymbol{Y} = (Y_1\ Y_2\ \cdots\ Y_k)^T \sim N(\boldsymbol{0}, \boldsymbol{C})$, where

$$\boldsymbol{C} = \begin{pmatrix} 1-p_1 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} & \cdots & -\sqrt{p_1p_k} \\ -\sqrt{p_2p_1} & 1-p_2 & -\sqrt{p_2p_3} & \cdots & -\sqrt{p_2p_k} \\ -\sqrt{p_3p_1} & -\sqrt{p_3p_2} & 1-p_3 & \cdots & -\sqrt{p_3p_k} \\ \vdots & \vdots & & \ddots & \vdots \\ -\sqrt{p_kp_1} & -\sqrt{p_kp_2} & -\sqrt{p_kp_3} & \cdots & 1-p_k \end{pmatrix}.$$

   It is stated in the proof that it now follows that $\boldsymbol{Y}^T\boldsymbol{Y} \sim \chi^2(k-1)$. Prove this. (2p)

Have a nice weekend!

Solutions

TAMS24/TEN1 2019-08-23

1. (a) We know that

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

so since σ is known, we could use $\overline{X}$ directly. However, we might as well follow the usual procedure. If H_0 holds, then

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1).$$

The critical region C is chosen so that

$$P(Z \in C \mid H_0) = \alpha,$$

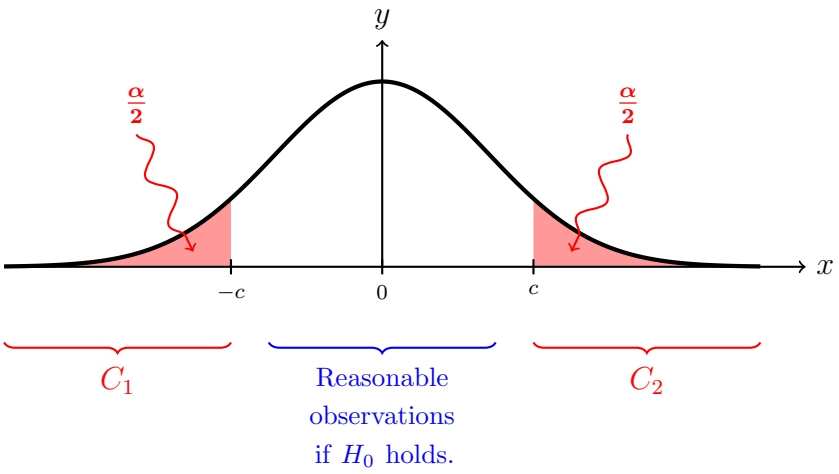
where we due to symmetry assume the form

$$C = \{z \in \mathbf{R} : |z| > c\} = \{z \in \mathbf{R} : z > c \text{ or } z < -c\}.$$

We note that C consists of two parts C_1 and C_2, where C_1 is on the negative half-axis. Due to symmetry,

$$P(Z \in C_1) = P(Z \in C_2) = \frac{\alpha}{2}.$$

In our case, $\alpha = 0.05$, so $c = \Phi^{-1}(0.975) = 1.96$.



Using our observations, we find that the test statistic is given by

$$z = \frac{\overline{x} - 3.0}{\sqrt{0.0625}/\sqrt{8}} = 1.7118 \notin C$$

so we can not reject H_0.

(b) The power at $\mu = 3.1$ can be calculated straight from the definition. Remember that $Z = \dfrac{\overline{X} - 3.0}{\sqrt{0.0625}/\sqrt{8}}$, so if $\mu = 3.1$, then

$$Z \sim N\left(\frac{0.1}{\sqrt{0.0625}/\sqrt{8}}, 1\right) = N(1.1314, 1).$$

Hence,

$$h(3.1) = P(H_0 \text{ rejected} \,|\, \mu = 3.1) = P(Z \in C \,|\, \mu = 3.1) = P\left(Z < -1.96 \text{ eller } Z > 1.96\right)$$
$$= \Phi(-1.96 - 1.1314) + 1 - \Phi(1.96 - 1.1314) = 0.2047.$$

(c) Given the observation $\overline{x} = 3.1513$ and $z = 1.7118$, we can follow the same procedure as in the previous exercise. However, in this case with C unknown. It is clear that we want to choose $c = 1.7118$ and since we found $c = \Phi^{-1}(1 - \alpha/2)$, it follows that

$$1.7118 = \Phi^{-1}(1 - \alpha/2) \quad \Leftrightarrow \quad \Phi(1.7118) = 1 - \frac{\alpha}{2},$$

so $\alpha = 2(1 - \Phi(1.7118)) = 0.087$. So we can choose the significance level 8.7% and still reject H_0. This is not good in practice! You should not look at the data to choose your significance level.

Answer: (a) We can not reject H_0. (b) 0.205 (c) $\alpha = 0.087$.

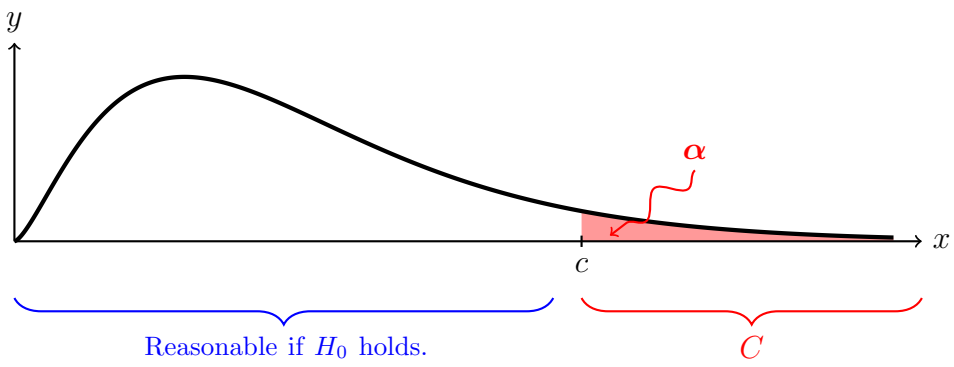
2. (a) Let $H_0 : \sigma^2 = 0.0625$ and the alternate hypothesis be $H_1 : \sigma^2 > 0.0625$. We have $n = 8$ samples, so

$$V = \frac{7\,S^2}{0.0625} \sim \chi^2(7).$$

We're seeking a limit c so that

$$\alpha = P(V > c) = P\left(S^2 > \frac{c\sigma_0^2}{n-1}\right)$$

and define the critical region as $C =]c, \infty[$.



From a table, we find $c = 14.07$, so

$$v = \frac{7 \cdot s^2}{0.0625} = 17.39 \in C.$$

We therefore reject H_0 and claim that it is very likely that the variance is greater than 0.0625.

(b) It follows that (by Cochran's and Gosset's theorems)

$$T_X = \frac{\overline{X} - \mu_X}{S/\sqrt{8}} \sim t(7),$$

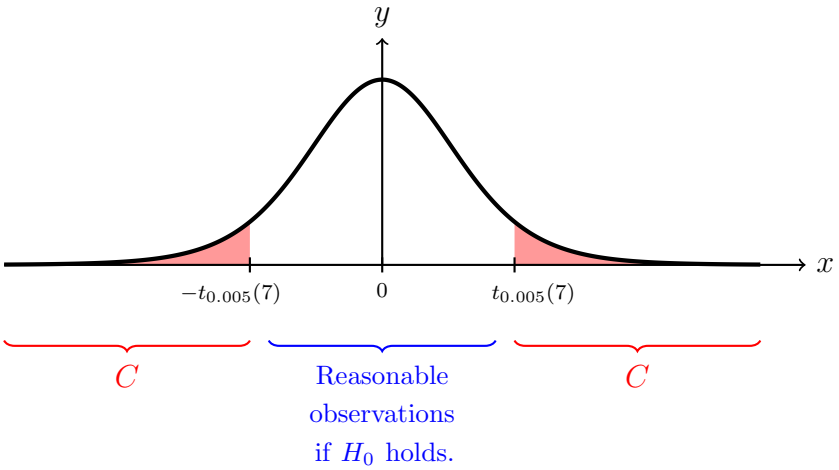
and

$$P(-t_{\alpha/2}(7) < T_X < t_{\alpha/2}(7)) = 1 - \alpha,$$

where we can solve the inequality for

$$\overline{X} - t_{\alpha/2}(7) \cdot \frac{S}{\sqrt{8}} < \mu_X < \overline{X} + t_{\alpha/2}(7) \cdot \frac{S}{\sqrt{8}}.$$

From a table, we find that $t_{0.005}(7) = 3.50$.



As an observation of S_X, we use $\sqrt{s_X^2}$, so

$$t_{0.005}(7)\frac{s}{\sqrt{8}} = 3.50 \cdot \frac{0.3943}{2.8284} = 0.4879.$$

Since $\overline{x} = 3.1513$, the interval is given by

$$I_{\mu_X} = (2.66,\ 3.64).$$

Answer: (a) We reject H_0. We believe that the variance is higher. (b) (2.66, 3.64).

3. We need to organize the data so that we can see how many machines were taken out of action in each time interval. We also need to choose these intervals so that — under the assumption that times are Exp($\mu = 1.0$)-distributed — all intervals are comparable in probability. The rule of thumb is to use $50/10 = 5$ classes, so we can try the following.

Time	How many
$I_1 = [0,\ 6)$	11
$I_2 = [6,\ 12)$	8
$I_3 = [12,\ 24)$	12
$I_4 = [24,\ 36)$	8
$I_5 = [36,\ \infty)$	11

It is not clear that this partitioning is good enough. Let H_0 be the hypothesis that times are Exp($\mu = 1.0$)-distributed. If H_0 is true, then the probability density of the time before a machine has to have its blades sharpened is given by $f(x) = \mu^{-1}\exp(-\mu^{-1}\,x)$ (with $\mu = 1.0$), so

$$P(a \le X < b) = \int_a^b \frac{1}{\mu}\exp\left(-\frac{x}{\mu}\right)dx = \exp\left(-\frac{a}{\mu}\right) - \exp\left(-\frac{b}{\mu}\right).$$

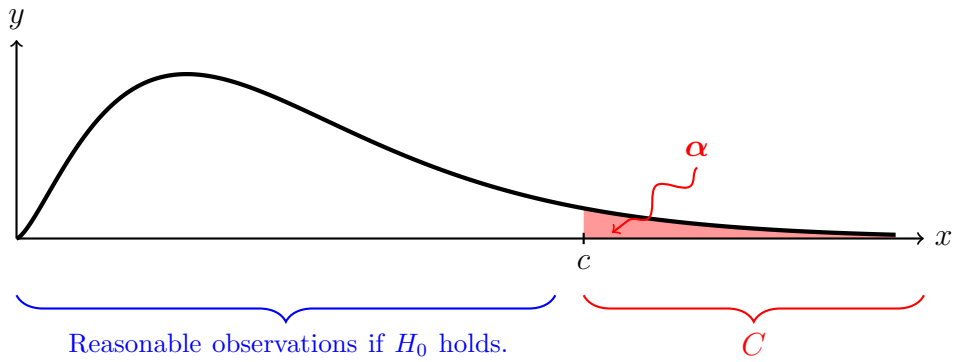
With these numbers, we can do the calculations and see that the probability of ending up in each interval is given by

$$P(X \in I_k) = \begin{cases} p_1 = 0.2212, & k = 1, \\ p_2 = 0.1723, & k = 2, \\ p_3 = 0.2387, & k = 3, \\ p_4 = 0.1447, & k = 4, \\ p_5 = 0.2231, & k = 5. \end{cases}$$

The testing quantity we now use is given by

$$q = \sum_{j=1}^{5} \frac{(x_j - np_j)^2}{np_j} = \frac{(11 - 50 \cdot 0.2212)^2}{50 \cdot 0.2212} + \cdots + \frac{(11 - 50 \cdot 0.2231)^2}{50 \cdot 0.2231} = 8.65.$$

If H_0 is true, then q is an observation of $Q \overset{\text{appr.}}{\sim} \chi^2(5-1) = \chi^2(4)$. We reject H_0 if q is large, so we need a critical region C of the form $C = [c, \infty)$. From a table we find that $c = \chi^2_{0.10}(4) = 7.78$. If $q \geq c$, we reject H_0.



Since $q \in C$, the conclusion is that we reject H_0. We do not believe that *Pleasures of the flesh* is telling the truth.

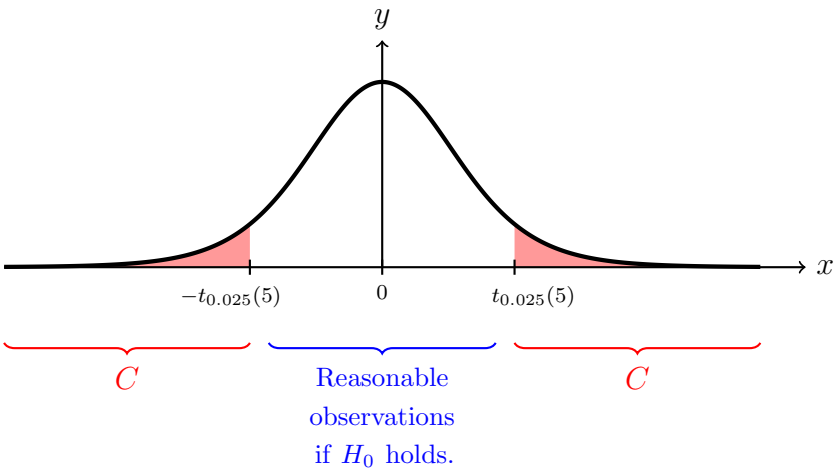
Answer: We reject the assumption.

4. (a) We can perform this test in several different ways. We can test whether $\beta_2 = 0$ in model 2 directly or we can compare model 1 and model 2 and see if model 2 is significantly better.

Alternative 1. To test if $\beta_2 = 0$, let $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 \neq 0$. Assume that H_0 holds. Then

$$T = \frac{\widehat{\beta}_2 - 0}{S\sqrt{h_{22}}} \sim t(8-3) = t(5),$$

where the distribution is clear since H_0 holds. We need a critical region C such that $P(T \in C \,|\, H_0) = 0.05$ and since H_1 is double sided, we choose symmetrically.



We find $t_{\alpha/2}(5) = t_{0.025}(5) = 2.57061$ in a table. An observation of $S\sqrt{h_{22}}$ is given by the standard error $d(\widehat{\beta}_2)$ and thus we find that the observation

$$t = \frac{1.0324}{0.2876} = 3.59$$

does belong to the critical region. So we reject H_0. The coefficient β_2 is probably not zero.

Alternative 2.
We have model 1:

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

and model 2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon,$$

where $x_2 = x_1^2$. We can test if the second model is significantly better by testing whether $\beta_2 = 0$ in a slightly different way.
Let

$$H_0 : \beta_2 = 0,$$

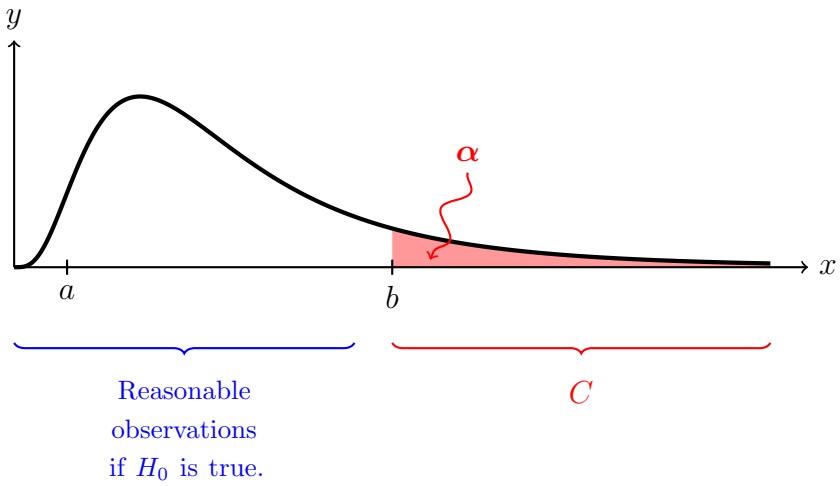
and

$$H_1 : \beta_2 \neq 0.$$

If H_0 is true, then $\boldsymbol{Y} \sim N(X_1\boldsymbol{\beta_1}, \sigma^2 I)$, so

$$W = \frac{(\mathrm{SS}_{\mathrm{E}}^{(1)} - \mathrm{SS}_{\mathrm{E}}^{(2)})/1}{\mathrm{SS}_{\mathrm{E}}^{(2)}/5} \sim F(1,5) \quad \text{if } H_0 \text{ is true}$$

since this is a quotient of independent χ^2 variables. If H_0 is not true, then W will tend to grow large. The critical region is given by $C =]c, \infty[$ for some $c > 0$.



From the table we find that $c = 10.0070$. An observation of W is found in

$$w = \frac{(0.0607 - 0.0170)/1}{0.0170/5} = 12.85,$$

so $w \in C$. We can reject the null hypothesis.

(b) We wish to find a confidence interval for β_1 using model 2. We know that

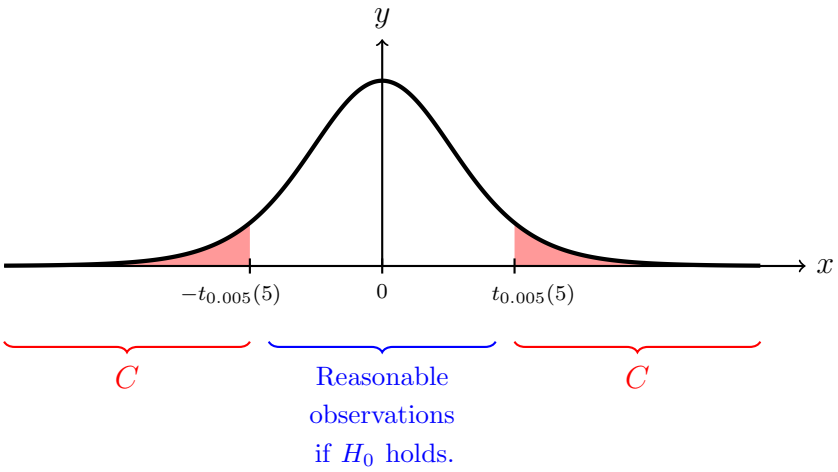
$$T = \frac{\widehat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(5).$$

So

$$P(-t_{\alpha/2}(5) < T < t_{\alpha/2}(5)) = 1 - \alpha,$$

where we can solve the inequality for

$$\widehat{\beta}_1 - t_{\alpha/2}(5) \cdot S\sqrt{h_{11}} < \beta_1 < \widehat{\beta}_1 + t_{\alpha/2}(5) \cdot S\sqrt{h_{11}}.$$



From a table, we find that $t_{0.005}(4) = 4.0321$. An observation of $S\sqrt{h_{11}}$ is given by the standard error $d(\widehat{\beta}_1) = 0.3315$ and thus we find the confidence interval

$$I_{\beta_1} = \left(\widehat{\beta}_1 - 4.0321 \cdot 0.3315,\ \widehat{\beta}_1 + 4.0321 \cdot 0.3315\right) = (-0.37,\ 2.30).$$

A similar calculation for model 1 yields

$$I_{\beta_1} = \left(\widehat{\beta}_1 - 4.0321 \cdot 0.1241,\ \widehat{\beta}_1 + 4.0321 \cdot 0.1241\right) = (1.62,\ 2.63).$$

Different intervals with different interpretations. For model 2, we could perform a hypothesis test to investigate whether $\beta_1 = 0$ or not and the conclusion would be that $\beta_1 = 0$ might very well be true. For model 1, the conclusion is that $\beta_1 \neq 0$.

Answer:

(a) A significance test shows that we conclude that $\beta_2 \neq 0$ at the significance level 5%.

(b) $(-0.37,\ 2.30)$. Different intervals with different interpretations. See above.

5. (a) Let X_1 be the number of fish eaten in the first tank and X_2 the number of fish eaten in the second tank. Assuming independence, it is clear that $X_1 \sim \text{Bin}(300, p_1)$ and that $X_2 \sim \text{Bin}(300, p_2)$. As estimators we choose

$$\widehat{P_1} = \frac{X_1}{300}, \quad \widehat{P_2} = \frac{X_2}{300}, \quad \text{and} \quad \widehat{P_2 - P_1} = \widehat{P_2} - \widehat{P_1}.$$

We note that $E(\widehat{P_1}) = E(X_1)/300 = 300p_1/300 = p_1$ and similarly $E(\widehat{P_2}) = p_2$, so $E(\widehat{P_2 - P_1}) = p_2 - p_1$. Our estimators are unbiased.
Now, we have observed that $\widehat{p_1} = 100/300 = 1/3$ and that $\widehat{p_2} = 120/300 = 2/5$. The binomial distribution is a bit messy to deal with in this instance (discrete intervals?), so let's try an approximation instead. Since

$$300 \cdot \widehat{p_1} \cdot (1 - \widehat{p_1}) = 300 \cdot \frac{1}{3} \cdot \frac{2}{3} = 66.67$$

and

$$300 \cdot \widehat{p_2} \cdot (1 - \widehat{p_2}) = 300 \cdot \frac{2}{5} \cdot \frac{3}{5} = 72$$

are both greater than 10, a normal approximation is reasonable. Hence,

$$\widehat{P_1} \overset{\text{appr.}}{\sim} \text{N}(p_1, p_1(1 - p_1)/300)$$

and

$$\widehat{P_2} \overset{\text{appr.}}{\sim} \text{N}(p_2, p_2(1 - p_2)/300).$$

From this it also follows that

$$\widehat{P_2} - \widehat{P_1} \overset{\text{appr.}}{\sim} \text{N}(p_2 - p_1, p_1(1 - p_1)/300) + p_2(1 - p_2)/300).$$

We now have

$$Z_1 = \frac{\widehat{P_1} - p_1}{\sqrt{\widehat{p_1}(1 - \widehat{p_1})/300}} \overset{\text{appr.}}{\sim} N(0, 1).$$

Note that we've replaced p_1 by the estimate $\widehat{p_1}$ in the denominator. Similarly

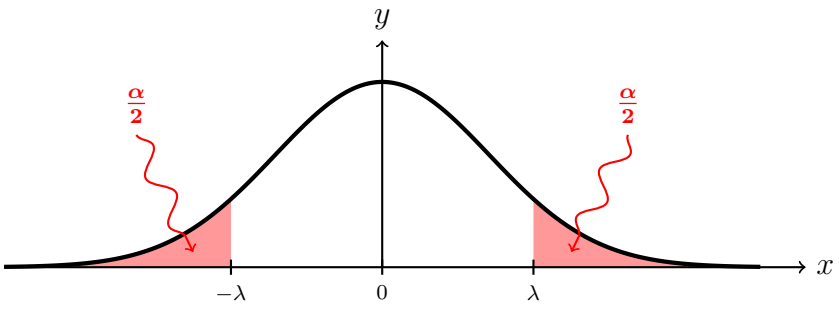
$$Z_2 = \frac{\widehat{P_2} - p_2}{\sqrt{\widehat{p_2}(1 - \widehat{p_2})/300}} \overset{\text{appr.}}{\sim} N(0, 1)$$

and

$$Z = \frac{\widehat{P_2} - \widehat{P_1} - (p_2 - p_1)}{\sqrt{\widehat{p_2}(1 - \widehat{p_2})/300 + \widehat{p_1}(1 - \widehat{p_1})/300}} \overset{\text{appr.}}{\sim} N(0, 1).$$

One reason for approximating the denominator is that we can use the normal distribution directly (with known variance). We seek a number λ so that, e.g.,

$$P(-\lambda < Z < \lambda) = 0.95.$$



We find $\lambda = 1.96$ from a table ($\lambda = \Phi^{-1}(0.975)$). So approximate confidence intervals (with 95% degree of confidence) can be found in

$$I_{p_1} = (0.333 - 1.96 \cdot 0.0272,\ 0.333 + 1.96 \cdot 0.0272) = (0.28, 0.39),$$

$$I_{p_2} = (0.4 - 1.96 \cdot 0.0283,\ 0.4 + 1.96 \cdot 0.0283) = (0.34, 0.46),$$

and since

$$\sqrt{\widehat{p_1}(1-\widehat{p_1})/300 + \widehat{p_2}(1-\widehat{p_2})/300} = 0.0393,$$

we obtain

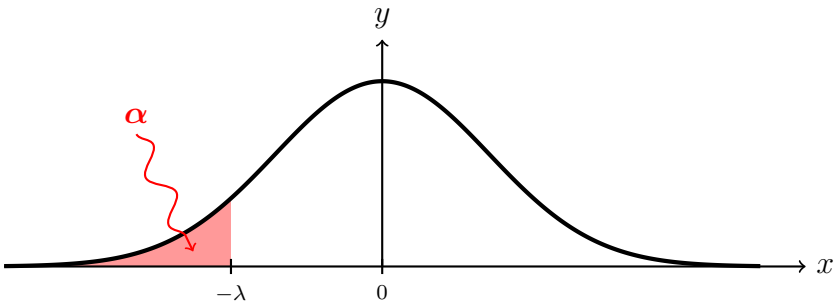
$$I_{p_2-p_1} = (0.0667 - 1.96 \cdot 0.0393,\ 0.0667 + 1.96 \cdot 0.0393) = (-0.01, 0.15).$$

Since $0 \in I_{p_2-p_1}$, we can't reject the hypothesis that $p_1 = p_2$.

(b) In this case, we want to test against the alternate hypothesis that $p_2 > p_1$ (not that $p_1 \neq p_2$). We use the same significance level, but place all the uncertainty in one tail. We use the same estimator for $p_2 - p_1$ and transform according to

$$Z = \frac{\widehat{P_2} - \widehat{P_1} - (p_2 - p_1)}{\sqrt{\widehat{p_2}(1-\widehat{p_2})/300 + \widehat{p_1}(1-\widehat{p_1})/300}} \overset{\text{appr.}}{\sim} N(0,1).$$

We seek λ so that $P(Z > -\lambda)$.



Since all the probability α is in the left tail, this pushes λ closer to the origin. From a table we find that $\lambda = \Phi^{-1}(0.95) = 1.645$. Then

$$-\lambda < Z \quad \Leftrightarrow \quad \widehat{P_2} - \widehat{P_1} - \lambda\sqrt{\widehat{p_2}(1-\widehat{p_2})/300 + \widehat{p_1}(1-\widehat{p_1})/300} < p_2 - p_1.$$

Using the point estimates for $\widehat{P_1}$ and $\widehat{P_2}$, we find that

$$p_2 - p_1 > 0.0021.$$

Hence $I_{p_2-p_1} = (0.0021, 1)$. We can now reject the hypothesis that $p_2 = p_1$ and claim that $p_2 > p_1$ is very likely.

Answer: (a) $I_{p_1} = (0.28, 0.39)$, $I_{p_2} = (0.34, 0.46)$ and $I_{p_2-p_1} = (-0.01, 0.15)$. Nope. (b) $I_{p_2-p_1} = (0.0021, 1)$. The second type is likely more ferocious.

6. We note that the covariance matrix can be written more compactly as $\boldsymbol{C} = I - \boldsymbol{q}\boldsymbol{q}^T$, where $\boldsymbol{q} = (\sqrt{p_1}\ \sqrt{p_2}\ \cdots\ \sqrt{p_k})^T$. Using this representation, we can verify that

$$(I - \boldsymbol{q}\boldsymbol{q}^T)^2 = I - \boldsymbol{q}\boldsymbol{q}^T \quad \text{and} \quad (I - \boldsymbol{q}\boldsymbol{q}^T)^T = I - \boldsymbol{q}\boldsymbol{q}^T,$$

so $\boldsymbol{C} = I - \boldsymbol{q}\boldsymbol{q}^T$ is a projection matrix and therefore has the eigenvalues $\lambda = 0$ and $\lambda = 1$. For these types of matrices, we know that the rank is equal to the trace. Since the trace of the matrix is equal to the sum of the eigenvalues, it is clear that

$$\operatorname{rank}(I - \boldsymbol{p}\boldsymbol{p}^T) = \operatorname{tr}(I - \boldsymbol{p}\boldsymbol{p}^T) = k - (p_1 + p_2 + \cdots + p_n) = k - 1,$$

so $\lambda = 0$ is a simple eigenvalue. The matrix is symmetric and positive semidefinite, so there exists an orthonormal matrix U such that $U^T\boldsymbol{C}U = \operatorname{diag}(1, 1, \ldots, 1, 0)$ becomes a diagonal matrix. If we let $\boldsymbol{Z} = U\boldsymbol{Y}$, we see that $\boldsymbol{Z} \sim N(\boldsymbol{0}, \operatorname{diag}(1, 1, \ldots, 1, 0))$ and that

$$\boldsymbol{Y}^T\boldsymbol{Y} = (U^T\boldsymbol{Z})^T U^T\boldsymbol{Z} = \boldsymbol{Z}^T\boldsymbol{Z} = \sum_{j=1}^{k-1} Z_j^2,$$

where $Z_j \sim N(0, 1)$ are independent. This sum is obviously $\chi^2(k-1)$-distributed!

Answer: See above.

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk statistik

Utbildningskod: TAMS24
Modul: TEN1
Datum: 2019-09-07
Institution: MAI

Exam in Statistics

TAMS24/TEN1 2019-09-07

You are permitted to bring:

- a calculator (no computer);
- Formel- och tabellsamling i matematisk statistik (from MAI);
- Formel- och tabellsamling i matematisk statistik, TAMS65;
- TAMS24: Notations and Formulas (by Xiangfeng Yang).

Grading (sufficient limits): 8-11 points giving grade 3; 11.5-14.5 points giving grade 4; 15-18 points giving grade 5. Your solutions need to be complete, well motivated, carefully written and concluded by a clear answer. Be careful to show what is random and what is not. Assumptions you make need to be explicit. Approximations are allowed if reasonable and clearly motivated. The exercises are in random order.

Solutions can be found on the homepage a couple of hours after the finished exam.

Once upon a time, in a land far far away, Rick and Gary had run into problems with their flamingo farm Exodus. Someone had posted threatening notes all around the farm. Gary suggested that it was the animal liberation army (or was it perhaps the different organization known as the liberation army of the animals?) that disliked the fact that there was flamingos in captivity (and, perhaps, also the slaughtering of said animals). Rick — on the other hand — assumed that it was the cult (devoted to the crawling chaos) that usually drifted around in the wooden area mumbling on about the great old ones and the unmentionable horrors at the mountains of madness. In either case, the result was that Exodus was shut down and Gary and Rick went into hiding. When they finally managed to return, things had taken a turn for the worse. The beautiful exodus sign had been scribbled over with frantic writing in something brownish red, stating that *Beneath the Columns of Abandoned Gods lies Dormant Hallucinations, where the Conjuration of the Sepulchral results in The Sleep of Morbid Dreams. In The Dead of Winter, Pestilential Winds causes the Exhumation of the Ancient.*
The air felt stale and suddenly there was no wind. Slowly, they entered the compound.

1. Since the flamingos had been left to their own devices while Exodus was shut down, they had managed to get into the storage where all the growth hormone was kept. By some coincidence — or perhaps supernatural reason caused by colors out of space — some of the flamingos had managed to ingest huge amounts of hormone and developed rapidly. Taking a random sample of the surviving flamingos (apparently aggression levels had gone up causing quite a lot of conflict), Gary wants to investigate the current state of affairs. Gary's measurements can be seen below.

   $$7.18 \quad 7.69 \quad 6.09 \quad 7.31 \quad 7.09 \quad 6.46 \quad 6.80 \quad 7.10$$

   Assume that the samples are independent and from a normal distribution with unknown variance σ^2 and an unknown expectation μ_{now}.

   (a) Find a confidence interval (99% degree of confidence) for the expectation μ_{now}. (1p)

   (b) Gary also found his old notes from before where he obtained the following random sample of weights.

   $$2.69 \quad 2.90 \quad 3.23 \quad 3.52 \quad 2.65 \quad 3.71 \quad 3.46 \quad 3.05$$

   Assume that these two samples are independent and that the old ones are from an $N(\mu_{\text{old}}, \sigma^2)$-distribution. Test the hypothesis that the expected weight before is less than half of the current expected weight at the significance level 5%. (2p)

   (c) Test the assumption that the variance of the two previous samples is the same. Use the significance level 5%. Conclusion? (2p)

2. After Gary shared his findings, Rick came clean about a mistake he made before the shut down. A New Mexico-native woman called Trinity had sold him some beautiful sand, full of greenish glass-like particles, that she had brought with her from the desert near Alamogordo. Rick had poured out several hundreds of kilos all around the water pond where the flamingos congregated. Unfortunately, it turned out that the sand contained high amounts of plutonium and fission products thereof. Rick is worried that the radioactivity is affecting the flamingos, so he takes some measurements using his old but trustworthy Geiger counter. The counter is calibrated to measure Giga Becquerel (1 Bq means 1 decay per second). Assume that the decay can be characterized by a Poisson process $X(t)$ with intensity $\lambda > 0$. Assume that Rick's counter reading is an observation of an $X = X(1) \sim \text{Po}(\mu)$ variable. When Rick took his measurement he obtained $x = 8$. For some reason, Rick felt that if the expectation μ wasn't greater than 5, everything was good enough (Not great. Not terrible).

   Let $H_0 : \mu = 5$ and $H_1 : \mu > 5$.

   (a) Perform a hypothesis test using the null hypothesis H_0 and the alternate hypothesis H_1 at the significance level 5%. (2p)

   (b) What is the power of the test at $\mu = 10$ (1p)

3. Assuming the same situation as in the previous exercise, additionally assume that we measure for 10 seconds and obtain the observation $y = 82$ of the random variable $Y = X(10)$ (meaning that we measure the stochastic process $X(t)$ for 10 seconds; $t = 10$). Find a 90% confidence interval for the intensity λ. (2p)

4. During their absence, cult members had obviously gained access to the farm and carried out their vile and unspeakable rituals. Both Gary and Rick found that the place felt very different from before. Shadows were angling in weird ways and from distant and terrible dimensions, echoes could be heard: *"Cthulhu fhtagn! Cthulhu fhtagn! Iä! Shub-Niggurath! The Goat with a Thousand Young!"*

   Not only had the flamingos grown a lot larger, but the radioactive sludge that had been formed in the water seemed to combine with the abysmal incantations, causing some of the flamingos to mutate and start forming tentacles. When using resonance amplifiers to increase the power of the echo from beyond, the tentaclification seemed to depend both on the intensity of the echoes and the level of radiation observed. It was unclear though, if administered growth hormone had any effect on the tentacles. To answer the obvious questions, a model was proposed:

$$\text{Model:} \quad Y = \beta_0 + \beta_1 a + \beta_2 r + \beta_3 h + \epsilon,$$

   where $\epsilon \sim N(0, \sigma^2)$ and different measurements are assumed to be independent. The quantity a is amplifier power and r is the radioactive intensity (in suitable units). The growth hormone is a binary where $h = 1$ means that growth hormone has been administered. The variables and measurements can be found below.

y	a	r	h
59.72	10	9	0
10.70	0	9	1
21.95	4	3	0
28.65	4	8	0
48.18	8	8	1
41.42	8	2	1
12.51	2	1	1
32.28	5	6	0
33.03	4	12	1
33.67	6	4	0

   The abstract unit used to describe how tentaclified the flamingos had become was denoted tentacliness. The following calculations has already been carried out.

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	0.41	0.82
1	4.86	0.10
2	1.11	0.08
3	0.28	0.56

Analysis of variance		
	Degrees of freedom	Square sum
REGR	3	2063.96
RES	6	4.35
TOT	9	2068.32

$$(X^T X)^{-1} = \begin{pmatrix} 931 & -74 & -50 & -284 \\ -74 & 13 & 0 & 19 \\ -50 & 0 & 9 & -4 \\ -284 & 19 & -4 & 428 \end{pmatrix} \cdot 10^{-3}.$$

   (a) Find a prediction interval for Y, with 90% degree of confidence, when $a = 2$, $r = 5$ and $h = 0$. (2p)

   (b) Test the hypothesis that the addition of growth hormone has an effect at the significance level 1%. (1p)

5. The effect on the flamingos seemed to be the at its worst at a specific place near the edge of the water, where moving a flamingo was impossible due to twisting displays of noneuclidian geometry and nauseating vortexes swirling like maelstroms of bent light. Rick and Gary managed to trap 5 flamingos in a box without seeing them clearly. The question was how many of these flamingos that had been distinctly affected by the tentaclification process. To investigate the matter, the following procedure was carried out.

   Their friend Susan — who seemed to be less affected by everything — put her hand inside the box while keeping her eyes away to avoid madness. She then grabs hold of a random flamingo in the box and feels for tentacles. Then she immediately releases the flamingo (still inside the box). This process is repeated n times, each repetition independent of the previous ones. Let $X_i = 0$ if the flamingo in try i was unaffected and let $X_i = 1$ if it was tentaclified. Furthermore, let θ be the total number of affected flamingos in the box.

   (a) When carrying out the first 7 tries, the result was $x = 1, 0, 1, 0, 1, 1, 0$. Find the maximum likelihood estimation $\widehat{\theta}$ of θ in this instance. (2p)

   (b) Choose one of the following questions (and answer it). (1p)

      i. Prove that $\widehat{\Theta}$ is unbiased.

      ii. Prove that $\widehat{\Theta}$ is biased.

      iii. Argue for why the previous questions are difficult to answer.

6. Since it apparently was the day of the tentacle (DoTT), Rick and Gary wondered if the tentacle lengths on the different affected flamingos was independent. Rick pointed out that the lengths should be normally distributed and Gary thought that they might find the correlation between different lengths. Looking at some theorems, they find that uncorrelated normally distributed variables are independent!

   Let $X_1, X_2, \ldots, X_n \sim N(0, 1)$ be independent and let $\boldsymbol{A} \in \mathbf{R}^{n \times n}$ be invertible. Moreover, let $\boldsymbol{\mu} \in \mathbf{R}^n$. Show that the components in $\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{\mu}$, where $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, are independent if and only if the covariance matrix of $\boldsymbol{Y}$ is a diagonal matrix. (2p)

Ia! Ia! Cthulhu ftagnh!

Solutions

TAMS24/TEN1 2019-09-07

1. (a) Let $X_i \sim \mathrm{N}(\mu_X, \sigma^2)$ be the new random sample. It follows that (by Cochran's and Gosset's theorems)

$$T_X = \frac{\overline{X} - \mu_X}{S/\sqrt{8}} \sim t(7),$$

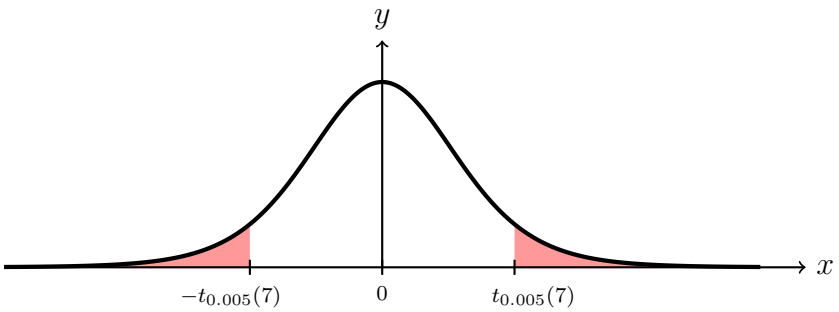
and

$$P(-t_{\alpha/2}(7) < T_X < t_{\alpha/2}(7)) = 1 - \alpha,$$

where we can solve the inequality for

$$\overline{X} - t_{\alpha/2}(7) \cdot \frac{S}{\sqrt{8}} < \mu_X < \overline{X} + t_{\alpha/2}(7) \cdot \frac{S}{\sqrt{8}}.$$

From a table, we find that $t_{0.005}(7) = 3.50$.



As an observation of S_X, we use $\sqrt{s_X^2}$, so

$$t_{0.005}(7)\frac{s}{\sqrt{8}} = 3.50 \cdot \frac{0.5032}{2.8284} = 0.6226.$$

Since $\overline{x} = 6.965$, the interval is given by

$$I_{\mu_X} = (6.34,\ 7.59).$$

(b) Let $Y \sim \mathrm{N}(\mu_Y, \sigma^2)$ be the old sample. Since the variances are equal, we weight them together according to the pooled variance:

$$s^2 = \frac{7s_1^2 + 7s_2^2}{14} = \frac{1}{2}\left(s_1^2 + s_2^2\right).$$

It now follows that (by Cochran's and Gosset's theorems)

$$T = \frac{0.5\overline{X} - \overline{Y} - (0.5\mu_X - \mu_Y)}{S\sqrt{\dfrac{0.5^2}{8} + \dfrac{(-1)^2}{8}}} \sim t(16-2) = t(14),$$

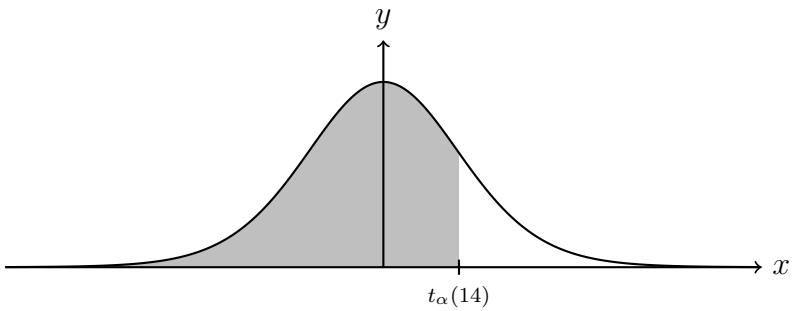
and

$$P(T < t_\alpha(14)) = 1 - \alpha,$$

where we can solve the inequality for

$$0.5 \cdot \overline{X} - \overline{Y} - t_\alpha(14) \cdot S\sqrt{\frac{0.5^2}{8} + \frac{1}{8}} < 0.5 \cdot \mu_X - \mu_Y.$$

We use a one-sided interval since we only want to investigate if $0.5 \cdot \mu_X > \mu_Y$. From a table, we find that $t_{0.05}(14) = 1.7613$.



As an observation of S, we use $\sqrt{s^2}$, so

$$t_{0.05}(14) \cdot s \cdot \sqrt{\frac{0.5^2}{8} + \frac{1}{8}} = 1.7613 \cdot 0.4520 \cdot 0.3953 = 0.3147.$$

Since $0.5 \cdot \overline{x} - \overline{y} = 0.3312$, the interval is given by

$$\begin{aligned} I_{0.5 \cdot \mu_X - \mu_Y} &= (0.3312 - 0.3147, \infty) \\ &= (0.0165, \infty). \end{aligned}$$

We see that 0 is not included in the interval, so we can claim that it is likely that $0.5 \cdot \mu_X > \mu_Y$ at this significance level.

(c) Let

$$H_0 : \sigma_X^2 = \sigma_Y^2 = \sigma^2$$

and

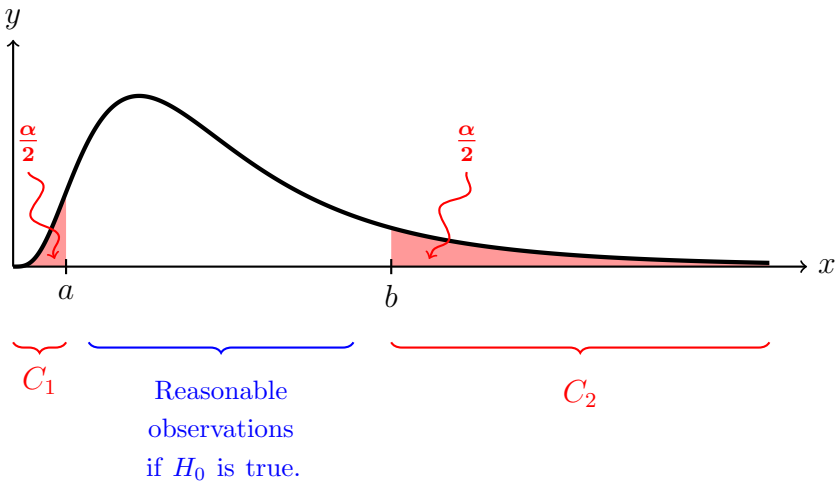
$$H_1 : \sigma_X^2 \neq \sigma_Y^2.$$

If H_0 is true, then $7S_X^2/\sigma^2 \sim \chi^2(7)$ and $7S_Y^2/\sigma^2 \sim \chi^2(7)$. Thus

$$V = \frac{\frac{7S_X^2}{\sigma^2}/7}{\frac{7S_Y^2}{\sigma^2}/7} = \frac{S_X^2}{S_Y^2} \sim F(7,7)$$

since S_1^2 and S_2^2 are independent. We seek a critical region C such that

$$\alpha = P(V \in C \,|\, H_0).$$



We find the bounds a and b from a table so that

$$P(V < a) = P(V > b) = \frac{\alpha}{2}.$$

with $a = 0.2002$ and $b = 5.00$. Note that a two-sided interval is necessary here. Since

$$v = \frac{0.2532}{0.1553} = 1.6306 \notin C$$

we can't reject H_0. The variances could be equal (but are they?)

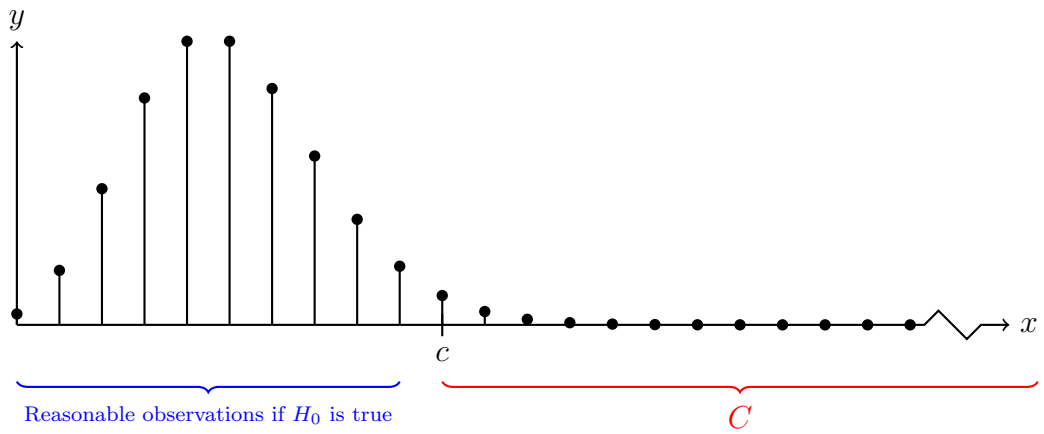
Answer:

(a) (6.34, 7.59)

(b) The new expectations seems to be more than twice the old one.

(c) We can't reject the hypothesis that they are equal; we do not know.

2. (a) Assume that H_0 is true. Let $X \sim \text{Po}(5)$ (since the expected number of counts is 5 during 1 second). We need to find the critical region C.



Let $p(k)$, $k = 0, 1, 2, \ldots$, be the probability function for X. From a table we can find that

$$\sum_{k=9}^{\infty} p(k) = 1 - \sum_{k=0}^{8} p(k) = 0.0681 \quad \text{and} \quad \sum_{k=10}^{\infty} p(k) = 0.0318.$$

Thus we choose

$$C = \{k \in \mathbf{Z} : k \geq 10\}.$$

Since our observation is $x = 8$ and $8 \notin C$, we can't reject H_0. It is possible that $\mu = 5$. Great news, right?!

(b) The power at $\mu = 10$ can be calculated straight from the definition:

$$\begin{aligned} h(10) &= P(H_0 \text{ rejected} \,|\, \mu = 10) = P(X \in C \,|\, \mu = 10) \\ &= \sum_{x=10}^{\infty} e^{-10} \frac{10^x}{x!} = 0.5421. \end{aligned}$$

Answer: (a) We can't reject H_0. It could be that $\mu = 5$ (we do not know). (b) 0.5421.

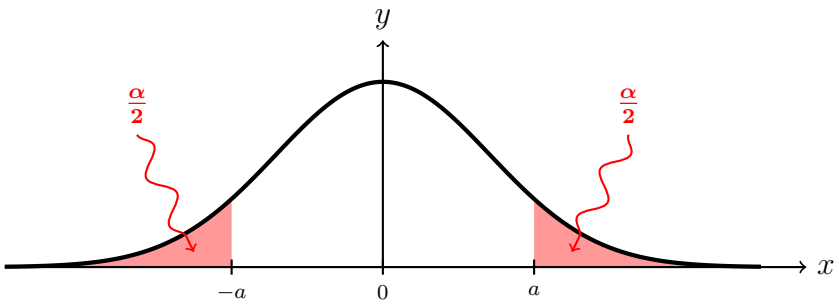
3. When measuring for 10 seconds, the expectation of $Y = X(10) \sim \text{Po}(10 \cdot \lambda)$ is $E(Y) = 10 \cdot \lambda$. We have observed $y = 82$, so it is reasonable to assume that $E(Y) > 15$. Thus we can use a normal approximation for Y. Moreover, $V(Y) = 10 \cdot \lambda$ (the Poisson distribution is funny..). Thus

$$Z = \frac{Y - 10\lambda}{\sqrt{10\widehat{\lambda}}} \overset{\text{appr.}}{\sim} N(0,1),$$

so we find $a > 0$ so that

$$0.90 = P(-a < Z < a).$$

Here we'll use the estimate $\widehat{\lambda} = 8.2$ (this will simplify matters).



So $a = 1.645$ is suitable. Then

$$-a < Z < a \quad \Leftrightarrow \quad -a < \frac{Y - 10\lambda}{\sqrt{10\widehat{\lambda}}} < a \quad \Leftrightarrow \quad \frac{Y - a\sqrt{10\widehat{\lambda}}}{10} < \lambda < \frac{Y + a\sqrt{10\widehat{\lambda}}}{10}$$

so with the observation $y = 82$ and estimate $\widehat{\lambda} = 8.2$, we obtain the (approximate) confidence interval

$$I_\lambda = (6.7, 9.7).$$

Answer: $I_\lambda = (6.7, 9.7)$.

4. (a) Let $u = (1\ 2\ 5\ 0)^T$ and let Y_0 be an independent random observation at $a = 2$, $r = 5$ and $h = 0$. Let $\widehat{\mu_0}$ be the estimate for the expectation μ at the same point. A well known test quantity is

$$T = \frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(10 - 4) = t(6).$$

We can box in this variable and solve for Y_0:

$$\begin{aligned} -t < T < t \quad &\Leftrightarrow \quad -t < \frac{Y_0 - \widehat{\mu_0}}{S\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} < t \\ &\Leftrightarrow \quad \widehat{\mu_0} - tS\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}} < Y_0 < \widehat{\mu_0} + tS\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}, \end{aligned}$$

where $t = t_{\alpha/2}(6) = t_{0.005}(6) = 1.9432$. We can now calculate that

$$\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u} = 0.412,$$

so $\sqrt{1 + \boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}} = 1.1883$. As an observation of S, we use

$$s = \sqrt{\frac{\text{SS}_\text{E}}{10 - 4}} = \sqrt{\frac{4.35}{6}} = \sqrt{0.725} = 0.851.$$

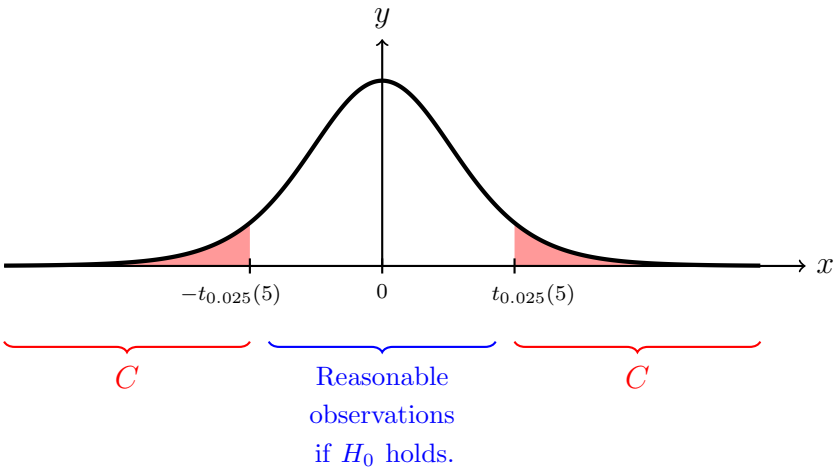
For $\widehat{\mu_0}$, we use the observation $\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} = 15.6755$. Thus we obtain the prediction interval

$$I_{Y_0} = \left(15.6755 \mp 1.9432 \cdot 0.851 \cdot 1.1883\right) = (13.7,\ 17.6).$$

(b) To test if $\beta_3 = 0$, let $H_0 : \beta_3 = 0$ and $H_1 : \beta_3 \neq 0$. Assume that H_0 holds. Then

$$T = \frac{\widehat{\beta_3} - 0}{S\sqrt{h_{33}}} \sim t(10-4) = t(6),$$

where the distribution is clear since H_0 holds. We need a critical region C such that $P(T \in C \,|\, H_0) = 0.01$ and since H_1 is double sided, we choose symmetrically.



We find $t_{\alpha/2}(6) = t_{0.005}(6) = 3.7074$ in a table. An observation of $S\sqrt{h_{33}}$ is given by the standard error $d(\widehat{\beta_3})$ and thus we find that the observation

$$t = \frac{0.28}{0.56} = 0.5$$

does not belong to the critical region. So we can't reject H_0. The coefficient β_2 might be zero.

Answer:

(a) (13.7, 17.6).

(b) A significance test shows that we can't conclude that $\beta_3 \neq 0$ at the significance level 1%. The addition of growth hormone might not have any effect.

5. (a) We start by noting that the parameter space is $\Omega_\theta = \{0, 1, 2, 3, 4, 5\}$. This means that continuous methods are problematic and it might be better to just find the ML-estimator directly. Why? Well, let us look at the likelihood function. Each X_i has the probability function

$$p_{X_i}(k) = \begin{cases} 1 - \frac{\theta}{5}, & k = 0, \\ \frac{\theta}{5}, & k = 1, \end{cases}$$

where the outcome $X_i = 1$ means that we've found a tentaclified flamingo. Thus the likelihood function is given by

$$L(\theta) = \prod_{i=1}^{n} p_{X_i}(x_k) = \left(1 - \frac{\theta}{5}\right)^{n - \sum x_k} \left(\frac{\theta}{5}\right)^{\sum x_k} = \left(1 - \frac{\theta}{5}\right)^{n(1-\overline{x})} \left(\frac{\theta}{5}\right)^{n\overline{x}}.$$

In our case we have $n = 7$, but we'll get to that. If we were to consider $L(\theta)$ as a function of a continuous $\theta > 0$, we could proceed as usual:

$$l(\theta) = \ln L(\theta) = n(1-\overline{x}) \ln \frac{5-\theta}{5} + n\overline{x} \ln \frac{\theta}{5},$$

so

$$0 = l'(\theta) = -\frac{n(1-\overline{x})}{5-\theta} + \frac{n\overline{x}}{\theta} \quad \Leftrightarrow \quad \theta = 5\overline{x}.$$

We can verify that $\widehat{\theta} = 5\overline{x}$ actually is a maximum of $l'(\theta)$ by observing the sign change of $l'(\theta)$, i.e., that we obtain $\nearrow$ max $\searrow$ for the function $l(\theta)$, so this would be our ML-estimate. However, there's no way to guarantee that $\widehat{\theta} \in \Omega_\theta$. We might surmise though that what we're looking for is close to $5\overline{x}$.
Using our sample $(x_1, \ldots, x_6) = (1, 0, 1, 0, 1, 1, 0)$, we see that

$$L(\theta) = \left(1 - \frac{\theta}{5}\right)^3 \left(\frac{\theta}{5}\right)^4,$$

so doing the calculations we find that

$$L(\theta) = 10^{-3} \cdot \begin{cases} 0, & \theta = 0, \\ 0.8192, & \theta = 1, \\ 5.5296, & \theta = 2, \\ 8.2944, & \theta = 3, \\ 3.2768, & \theta = 4, \\ 0, & \theta = 5. \end{cases}$$

We see that $\theta = 3$ provides the highest probability, so by definition this is the MLE. What happens with the continuous version? Well, we have $5\overline{x} = 5 \cdot \frac{4}{7} = 2.857$. Rounding it we'd obtain $\widehat{\theta} = 3$, but we would have to prove that this is the actual MLE in that case.

(b) This is a rather tricky question. If you'd try and just use the estimator $5\overline{x}$ (which gives values very likely outside of the parameter space), it's rather easy to show that it is unbiased:

$$E(5\overline{X}) = 5E(X) = 5\frac{\theta}{5},$$

where X has the same distribution as all the X_i:s in the average. But we can't really use this as our estimator. So what about the integer part of $5\overline{X}$? Well, the maximum should (considering the investigation above) happen if we either round up or down, so how would we know which one? Should we choose whichever is closest? What's to say that choosing in that way would yield the true MLE? So yeah.. difficult to answer :).

Answer: (a) $\widehat{\theta} = 3$ (b) see above.

6. Since A is invertible, we know that $\det A \neq 0$.

   $\Rightarrow$) This direction is more or less trivial. If the components of $\boldsymbol{Y}$ are independent, then $C(Y_i, Y_j) = 0$ for $i \neq j$ and $C(Y_i, Y_i) = \sigma_i^2$, so $C_{\boldsymbol{Y}}$ is obviously a diagonal matrix.

   $\Leftarrow$) Now suppose that $C_{\boldsymbol{Y}}$ is a diagonal matrix, e.g.,

$$\begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}.$$

   Since $C_{\boldsymbol{Y}} = AA^T$ we know that $C_{\boldsymbol{Y}}$ is invertible, which implies that $\sigma_i^2 \neq 0$ for all i. The inverse $C_{\boldsymbol{Y}}^{-1}$ is the diagonal matrix with the diagonal elements σ_i^{-2}. We thus obtain the joint density function

$$\begin{aligned} f_{\boldsymbol{Y}}(\boldsymbol{y}) &= \frac{1}{(2\pi)^{n/2}\sqrt{\det C_{\boldsymbol{Y}}}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T C^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right) \\ &= \frac{1}{(\sqrt{2\pi})^n \sigma_1\sigma_2\cdots\sigma_n} \exp\left(-\frac{1}{2}\sum_{j=1}^{n}(y_j-\mu_j)\sigma_j^{-2}(y_j-\mu_j)\right) \\ &= \prod_{j=1}^{n} \frac{1}{\sigma_j\sqrt{2\pi}} \exp\left(-\frac{(y_j-\mu_j)^2}{2\sigma_j^2}\right) = \prod_{j=1}^{n} f_{Y_j}(y_j). \end{aligned}$$

   We have now shown that the joint density function is given by the product of the density functions for Y_j, which immediately implies that the components of $\boldsymbol{Y}$ are independent.

Answer: See above.

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk statistik

Kurskod: TAMS24
Provkod: TEN1
Datum: 2018-10-44
Institution: MAI

Example exam in Statistics

TAMS24/TEN1 2018-10-44 8–12

You are permitted to bring:

- a calculator (no computer);
- Formel- och tabellsamling i matematisk statistik (from MAI);
- Formel- och tabellsamling i matematisk statistik, TAMS65;
- TAMS24: Notations and Formulas (by Xiangfeng Yang).

Grading: 8-11 points giving grade 3; 11.5-14.5 points giving grade 4; 15-18 points giving grade 5. Your solutions need to be complete, well motivated, carefully written and concluded by a clear answer. Be careful to show what is random and what is not. Assumptions you make need to be explicit. The exercises are in number order.

Solutions can be found on the homepage a couple of hours after the finished exam.

Exercises loosly based on an exam in TAMS65 by Martin Singull.

1. A study of use of cannabis amongst youth (445 people) and their parents use of alcohol and/or other narcotics gave the following table.

		Usage (youths)		
		Never	Sometimes	Regularly
Usage (parents)	None	141	54	40
	One	68	44	51
	Both	17	11	19

   Using a suitable test, examine if there is a connection between the parents use of alcohol and/or narcotics and the youth's use of cannabis. (2p)

2. A company has stores in 4 large cities and measure sales during 8 weeks. They divide the raw numbers by the population in each city to normalize and the result (rounded off) can be seen in the table below.

		Week									
		1	2	3	4	5	6	7	8	$\overline{x}$	s
City	1	19	14	12	17	21	16	15	19	16.50	2.88
	2	8	12	9	9	8	10	11	8	9.38	1.51
	3	13	6	6	8	11	8	9	9	8.75	2.38
	4	17	18	17	22	16	18	14	21	17.88	2.59

We assume that the data are observations of independent random variables such that on row i, we have observations of $N(\mu_i, \sigma^2)$ (the variance is σ^2). We assume that all rows have the same variance. Construct two-sided confidence intervals for $\mu_i - \mu_j$ such that the *simultaneous* degree of confidence is $\geq 94\%$. Are there differences between the cities? (3p)

3. For a number of coal driven power plants with similar construction, one has measured the level y of sulphur dioxide release (ppm) and the effect x (GW). A helpful technician has provided the following computer analysis of the data. The model used was

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ (the variance is σ^2) and different measurements are assumed to be independent. The regression line obtained was $y = 204 - 638x + 959x^2$ and we have the following data:

i	$\widehat{\beta}_i$	$d(\widehat{\beta}_i)$
0	204.46	82.95
1	-638.0	298.5
2	959.2	263.6

Analysis of variance

	Degrees of freedom	Square sum
REGR	2	15049.3
RES	6	254.3
TOT	8	15303.6

$$(X^T X)^{-1} = \begin{pmatrix} 162.38 & -581.90 & 508.70 \\ -581.90 & 2102.17 & -1850.58 \\ 508.70 & -1850.58 & 1639.75 \end{pmatrix}$$

(a) How many power plants were examined? (1p)

(b) Is the second degree term necessary? Motivate your answer at the level 5%. (1p)

(c) Find a 95% confidence interval for $E(Y)$ when the effect is 0.5 GW. (2p)

4. A company is measuring the quality of work and for a certain variable used 25 samples $x_1, \ldots, x_{25}$ are collected independently during a day. It is reasonable to assume that this sample comes from the distribution $N(\mu,\ \sigma^2 = 1.2^2)$ (where $\sigma^2 = 1.2^2$ is the variance). The company wants to have $\mu > 30$ and needs help performing the test $H_0 : \mu = 30$ against $H_1 : \mu > 30$.

(a) If $\overline{x} = 30.35$, carry out the test at the significance level 5%. (1.5p)

(b) For which values of μ is the power of the test at least 75%? (1.5p)

5. The life time of a certain type of electrical components has the probability density

$$f(x) = a^2 x e^{-ax}, \quad x \geq 0,$$

where $a > 0$ is an unknown constant. During a test-run, 50 such components were found to have a cumulative life time of 250 time units (the 50 life times added together).

(a) Find a reasonable point estimate for a. (2p)

(b) Find a 95% confidence interval – one-sided bounded from below –for the expected life time of such a component. (2p)

6. Prove that S^2 is an unbiased estimate for σ^2. (2p)

LINKÖPINGS UNIVERSITET
Matematiska institutionen
Matematisk Statistik

Kurskod: TAMS24
Provkod: TEN1
Datum: 2018-10-42
Institution: MAI

Example exam in Statistics (solutions)

1. We expand the table a bit to introduce some quantities.

		Usage (youth)			
		Never	Sometimes	Regularly	Sum
Usage (parents)	None	141	54	40	235
	One	68	44	51	163
	Both	17	11	19	47
Sum		226	109	110	445
$\widehat{p_j}$		0.5079	0.2449	0.2472	1.000

We see that $n_i\widehat{p_j} \geq 5$ for all 9 boxes, so we can perform a χ^2-test. Let

$$H_0 : \text{Parents use of alcohol/narcotics is independent of the adolescents use}$$

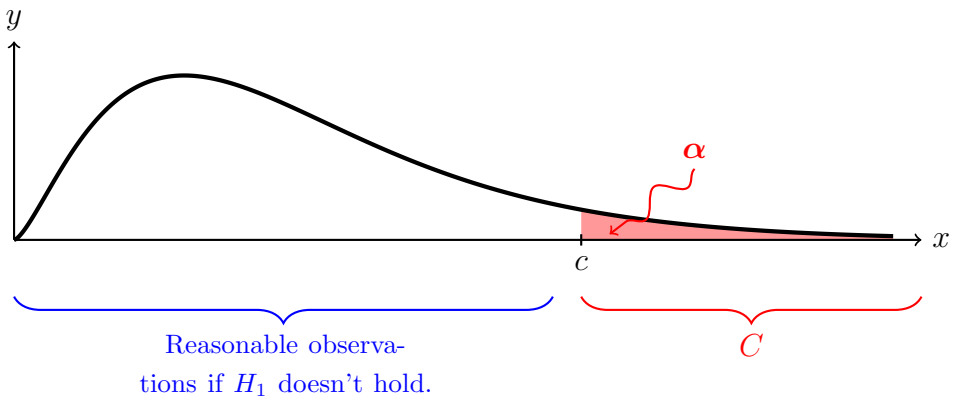
and

$$H_1 : \text{The usage is not independent.}$$

We calculate

$$q = \frac{(141 - 235 \cdot 0.5079)^2}{235 \cdot 0.5079} + \frac{(54 - 235 \cdot 0.2449)^2}{235 \cdot 0.2449} + \cdots + \frac{(19 - 47 \cdot 0.2472)^2}{47 \cdot 0.2472} = 22.3350.$$

If H_0 is true, then q is an observation of $Q \overset{\text{appr.}}{\sim} \chi^2((3-1)(3-1)) = \chi^2(4)$. We reject H_0 if q is large, so we need a critical region C. From a table we find that $c = \chi^2_{0.01}(4) = 13.28$ and we define $C = [c, \infty)$.



Since $q = 22.335 \in C$, we reject H_0. The adolescents use is not independent of the parents usage.

Answer: The adolescents use is not independent of the parents usage.

2. First, the requirement that the confidence intervals should be simultaneous needs to be addressed. What this means, is that if $I_1, I_2, \ldots, I_k$ are *independent* confidence intervals for $\theta_1, \theta_2, \ldots, \theta_k$, then

$$P(\theta_1 \in I_1 \text{ and } \theta_2 \in I_2 \ \cdots \ \text{ and } \theta_k \in I_k) = \prod_{j=1}^{k} P(\theta_j \in I_j) = (1-\alpha)^k,$$

if we use the same level of confidence $1-\alpha$ for all intervals. Here we used the fact that the variables used are independent, so that the confidence intervals are independent.

For this exercise, we need to compare four cities, so 6 confidence intervals are needed:

$$I_{\mu_1-\mu_2},\ I_{\mu_1-\mu_3},\ I_{\mu_1-\mu_4},\ I_{\mu_2-\mu_3},\ I_{\mu_2-\mu_4},\ I_{\mu_3-\mu_4}.$$

Is this clear? Well, we've assumed that all X_{ij} are independent, so the means $\overline{X}$ for each row are independent. By Cochran's theorem, we know that S^2 are independent of the mean for each line. Since the pooled variance is a function of the variances, it will also be independent of the means. So the limits of the confidence intervals are indeed independent of each other.

Now, by Bernoulli's inequality, we know that $(1-\alpha)^k \geq 1 - k\alpha$, so with $\alpha = 0.01$, the simultaneous degree of confidence will be better than 94% if we use 6 intervals. We could also just test different values for the confidence level in $(1-\alpha)^6$ until we hit a sweet spot.

We weight together the variances according to the pooled variance:

$$s^2 = \frac{7s_1^2 + 7s_2^2 + 7s_3^2 + 7s_4^2}{28} = \frac{1}{4}\left(s_1^2 + s_2^2 + s_3^2 + s_4^2\right).$$

For each row, let $\overline{X_i}$ denote the mean value of said row, $i = 1, 2, 3, 4$. It now follows that (by Cochran's and Gosset's theorems)

$$T = \frac{\overline{X_i} - \overline{X_j} - (\mu_i - \mu_j)}{S\sqrt{\frac{1}{8} + \frac{1}{8}}} \sim t(28),$$

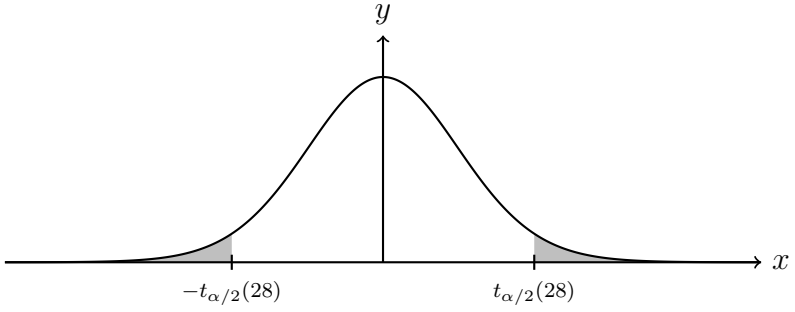
and

$$P(-t_{\alpha/2}(28) < T < t_{\alpha/2}(28)) = 1 - \alpha$$

where we can solve the inequality for

$$\overline{X_i} - \overline{X_j} - t_{\alpha/2}(28) \cdot \frac{S}{2} < \mu_i - \mu_j < \overline{X_i} - \overline{X_j} + t_{\alpha/2}(28) \cdot \frac{S}{2}.$$

Note that $\sqrt{\frac{1}{8} + \frac{1}{8}} = \frac{1}{2}$. From a table, we find that $t_{0.005}(28) = 2.7633$.



We use the observation $\sqrt{s^2} = \sqrt{5.74}$ of S, so

$$t_{0.005}(28)\frac{s}{2} = 3.3102.$$

We now get 6 interesting intervals for comparison of differences $\mu_i - \mu_j$. We use the observations $\overline{x_i} - \overline{x_j}$ of $\overline{X_i} - \overline{X_j}$ in each case, leading to

$$I_{\mu_i - \mu_j} = \left(\overline{x_i} - \overline{x_j} \mp 3.3102\right).$$

Thus

$$\begin{aligned} I_{\mu_1-\mu_2} &= (3.81,\ 10.43) \\ I_{\mu_1-\mu_3} &= (4.44,\ 11.06) \\ I_{\mu_1-\mu_4} &= (-4.69,\ 1.93) \\ I_{\mu_2-\mu_3} &= (-2.68,\ 3.94) \\ I_{\mu_2-\mu_4} &= (-11.81,\ -5.19) \\ I_{\mu_3-\mu_4} &= (-12.44,\ -5.82) \end{aligned}$$

We can see from this that $\mu_1 > \mu_{2,3}$ but that it is inconclusive if $\mu_1 \neq \mu_4$. Similarly, $\mu_2 < \mu_4$ and $\mu_3 < \mu_4$, but we do not know if $\mu_2 \neq \mu_3$ or not. The conclusion that can be drawn is that offices in cities 1 and 4 has better sales than 2 and 3, whereas nothing can be said about city 1 and 4 and between 2 and 3.

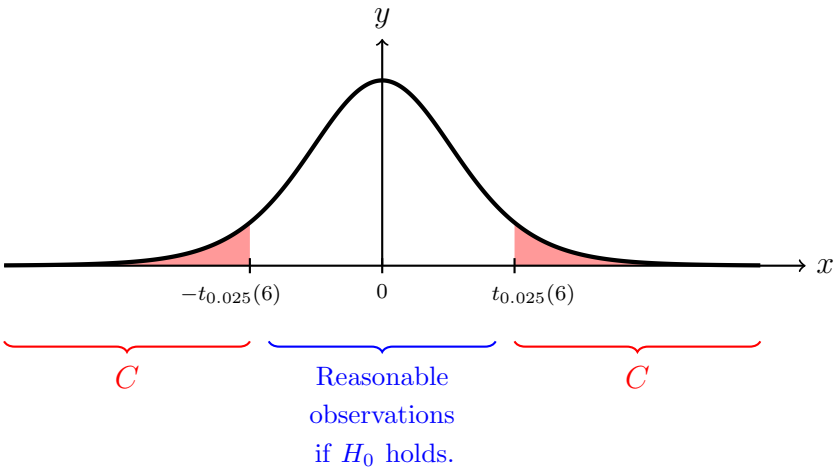
Answer: See above.

3. (a) 9 power plants were examined. This is clear since $\mathrm{SS_E}$ has $n-k-1$ degrees of freedom and $\mathrm{SS_R}$ has k. From the table we see that $n-k-1 = 6$ and $k = 2$, so $n = 9$.

   (b) To test if the second term is necessary, let $H_0 : \beta_2 = 0$ and $H_1 : \beta_2 \neq 0$. Assume that H_0 holds. Then

$$T = \frac{\widehat{\beta}_2 - 0}{S\sqrt{h_{22}}} \sim t(6),$$

   where the distribution is clear since H_0 holds. We need a critical region C such that $P(T \in C \,|\, H_0) = 0.05$ and since H_1 is double sided, we choose symmetrically.



   We find $t_{\alpha/2}(6) = t_{0.025}(6) = 2.4469$ from a table and we use the observation of $S\sqrt{h_{22}}$ in the form of the standard error $d(\widehat{\beta}_2)$. Thus we find that the observation

$$t = \frac{959.2}{263.6} = 3.64$$

   is in the critical region, so we reject H_0. The second degree term is useful.

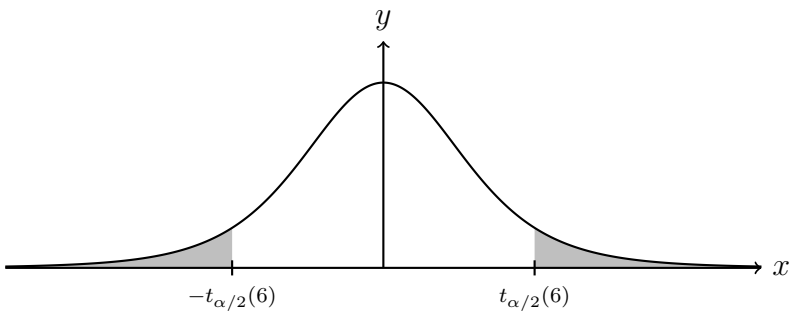
(c) To find a confidence interval for the expectation at $x = 0.5$, we let $u = (1\ 0.5\ 0.5^2)^T$. Let $\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}}$ be the estimate for μ at $x = 0.5$. Then (by Gosset's theorem)

$$T = \frac{\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - \boldsymbol{u}^T\boldsymbol{\beta}}{S\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}}} \sim t(6),$$

if we use $S^2 = \mathrm{SS_E}/6$. We box in T and solve for $\boldsymbol{u}^T\boldsymbol{\beta}$:

$$-t < T < t \quad \Leftrightarrow \quad \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} - tS\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}} < \boldsymbol{u}^T\boldsymbol{\beta} < \boldsymbol{u}^T\widehat{\boldsymbol{\beta}} + tS\sqrt{\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u}},$$

where $t = t_{\alpha/2}(6) = t_{0.025}(6) = 2.4469$.



A straight forward calculation yields $\boldsymbol{u}^T(X^TX)^{-1}\boldsymbol{u} = 0.2119$. As an observation of S, we use

$$s = \sqrt{\mathrm{SS_E}/6} = \sqrt{254.3/6} = 6.5102,$$

and $\boldsymbol{u}^T\widehat{\boldsymbol{\beta}} = 125.26$, from which we obtain the confidence interval

$$I_\mu = \left(125.26 \mp 2.4469 \cdot 6.5102 \cdot \sqrt{0.2119}\right) = (117.9,\ 132.6).$$

Answer: (a) $n = 9$ (b) It is useful. (c) $I = (117.9,\ 132.6)$.

4. (a) Assume that H_0 is true. Then $\overline{x}$ is an observation of

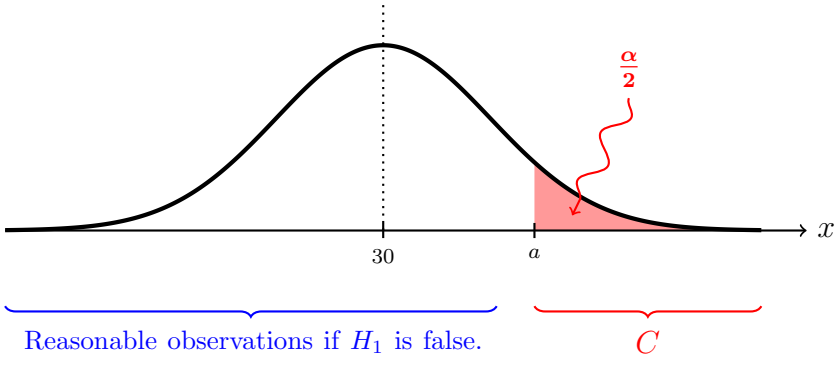
$$\overline{X} \sim N(30, 1.2^2/25) = N(30, 0.0576).$$

The critical region is of the form $C = [a, \infty)$ since H_1 is one sided and we're looking at a higher mean value. Since

$$0.05 = P(\overline{X} \geq a) = 1 - P(\overline{X} < a) = 1 - P\left(\frac{\overline{X} - 30}{0.24} < \frac{a - 30}{0.24}\right) = 1 - \Phi\left(\frac{a - 30}{0.24}\right),$$

we obtain that

$$\Phi^{-1}(0.95) = \frac{a - 30}{0.24} \quad \Leftrightarrow \quad a = 30 + 0.24 \cdot 1.645 = 30.3948.$$



Since $\overline{x} = 30.35$ was observed, we did not get an observation in the critical region. Hence we can not reject H_0. The expectation might very well be $\mu = 30$.

Note that it is unfeasible to use $\overline{x}$ as test statistic if σ is unknown (why?).

(b) The power at μ is defined as

$$h(\mu) = P(H_0 \text{ rejected} \,|\, \mu \text{ is the correct value}) = P(\overline{X} \geq a \,|\, \overline{X} \sim N(\mu, 0.24^2))$$
$$= 1 - \Phi\left(\frac{a-\mu}{0.24}\right).$$

If we want the power to be $\geq 75\%$, then

$$0.75 \leq 1 - \Phi\left(\frac{a-\mu}{0.24}\right) = \Phi\left(\frac{\mu-a}{0.24}\right) \quad \Leftrightarrow \quad 0.6745 \leq \frac{\mu-a}{0.24}$$
$$\Leftrightarrow \quad a + 0.6745 \cdot 0.24 \leq \mu,$$

since Φ is strictly increasing. Hence $\mu \geq 30.557$.

Answer: (a) We can't reject H_0 (b) $\mu \geq 30.557$.

5. We let X be a random variable with density function

$$f(x) = a^2 x e^{-ax}, \quad x \geq 0,$$

where $a > 0$ is an unknown constant.

(a) To find an estimate for a, let's use the method of moments (since it's usually the easiest). We find

$$E(X) = a^2 \int_0^\infty x^2 e^{-ax}\, dx = \cdots = \frac{2}{a},$$

using integration by parts for example. To find the MME, we solve for $\widehat{a}$ in

$$\frac{2}{\widehat{a}} = \overline{x} \quad \Leftrightarrow \quad \widehat{a} = \frac{2}{\overline{x}}$$

if $\overline{x} \neq 0$. We have observed that the cumulative lifetime of 50 units was 250, so we can estimate a by

$$\widehat{a} = \frac{2 \cdot 50}{250} = 0.4.$$

(b) By the CLT, we know that $\sum_{k=1}^{50} X_k \overset{\text{appr.}}{\sim} N(50\mu, 50\sigma^2)$, where σ^2 and μ are the variance and expectation of a single X, respectively. We know that $\mu = 2/a$ but need to calculate the variance. The second moment is given by

$$E(X^2) = a^2 \int_0^\infty x^3 e^{-ax}\, dx = \cdots = \frac{6}{a^2},$$

again using integration by parts. By Steiner's theorem, we thus obtain that

$$V(X) = E(X^2) - E(X)^2 = \frac{6}{a^2} - \frac{4}{a^2} = \frac{2}{a^2}.$$

Hence

$$Y := \sum_{k=1}^{50} X_k \overset{\text{appr.}}{\sim} N\left(\frac{100}{a}, \frac{100}{a^2}\right).$$

We are asked to find a confidence interval I_μ for μ, but since $\mu = 2/a$ we can start by finding one for a. Since I_μ should be bounded from below, we seek I_a bounded from above. Therefore we choose $c = \Phi^{-1}(0.95) = 1.645$ such that

$$0.95 = P\left(\frac{Y - 100/a}{10/a} \leq c\right).$$

Solving the inequality in the probability measure yields

$$Y - \frac{100}{a} \leq \frac{16.45}{a} \quad \Leftrightarrow \quad Y \leq \frac{116.45}{a} \quad \Leftrightarrow \quad a \leq \frac{116.45}{Y}.$$

Using the observation $y = 250$ of Y, we obtain a confidence interval $I_a = (0,\ 0.4658)$. Since $\mu = 2/a$, we therefore have $I_\mu = (4.29,\ \infty)$.

Answer: (a) $a = 0.4$ (b) $I_\mu = (4.29,\ \infty)$.

6. The sample variance is defined by $S^2 = \dfrac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Hence

$$E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2\right) = \frac{1}{n-1}E\left(\sum_{i=1}^{n} X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\big(E(X_i^2) - 2E(X_i\overline{X}) + E(\overline{X}^2)\big).$$

We know that $E(\overline{X}) = \mu$ and $V(\overline{X}) = \sigma^2/n$, and since $E(Y^2) = V(Y) + E(Y)^2$, we have

$$E(X_i^2) = V(X_i) + E(X_i)^2 = \sigma^2 + \mu^2 \quad \text{and} \quad E(\overline{X}^2) = \sigma^2/n + \mu^2.$$

Furthermore,

$$E(X_i\overline{X}) = E\left(X_i\frac{1}{n}\sum_{k=1}^{n} X_k\right) = \frac{1}{n}\sum_{k=1}^{n} E(X_iX_k)$$

and since $E(X_iX_k) = E(X_i)E(X_k) = \mu^2$ if $i \neq k$ (since these variables are independent) and $E(X_i^2) = \sigma^2 + \mu^2$ (when $i = k$), it follows that

$$E(X_i\overline{X}) = ((n-1)\mu^2 + \sigma^2 + \mu^2)/n = \mu^2 + \sigma^2/n.$$

In summary, we have now shown that

$$E(S^2) = \frac{1}{n-1}\sum_{i=1}^{n}\big(\sigma^2 + \mu^2 - 2(\mu^2 + \sigma^2/n) + \sigma^2/n + \mu^2\big) = \frac{n\sigma^2 - n\sigma^2/n}{n-1} = \sigma^2,$$

so S^2 is an unbiased estimator for σ^2.

Answer: See above.