Solutions

TAMS24/TEN1 2019-01-09

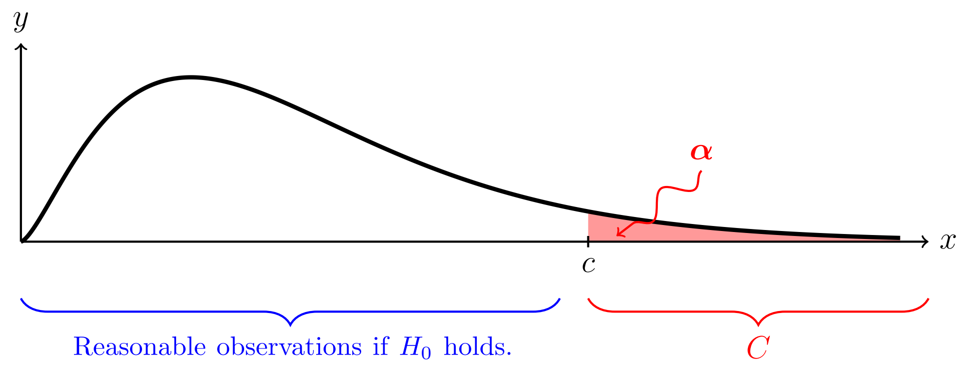
1. Let H_0 be the hypothesis that the data is homogeneous between the two sites and H_1 that this is not true. In total, we have $n = 289$ observations. We can directly see that the last four albums will have too small $n_i\widehat{p}_j$ (significantly less than 5), so we have to combine these to obtain 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

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

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)

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

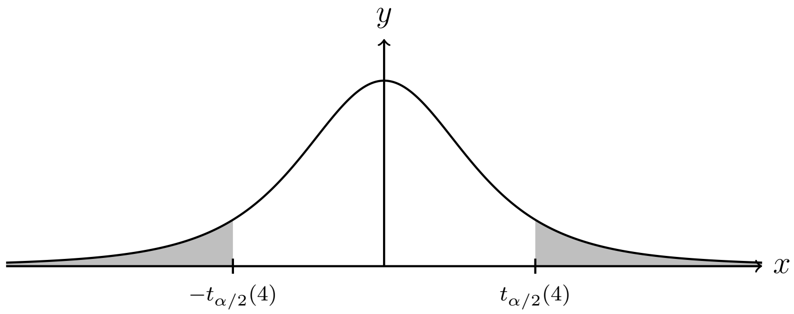
and

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

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,

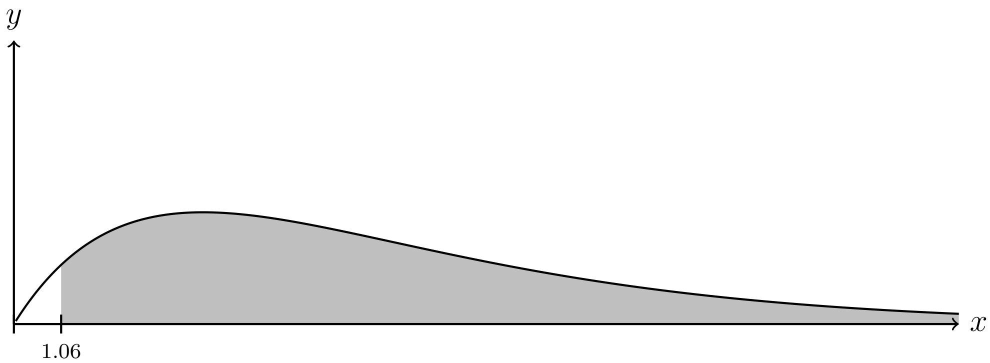
$$I_\Delta = (8.4 - 2.7764 \cdot 7.2319/\sqrt{5},\, 8.4 + 2.7764 \cdot 7.2319/\sqrt{5}) = (-0.58,\, 17.4).$$

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

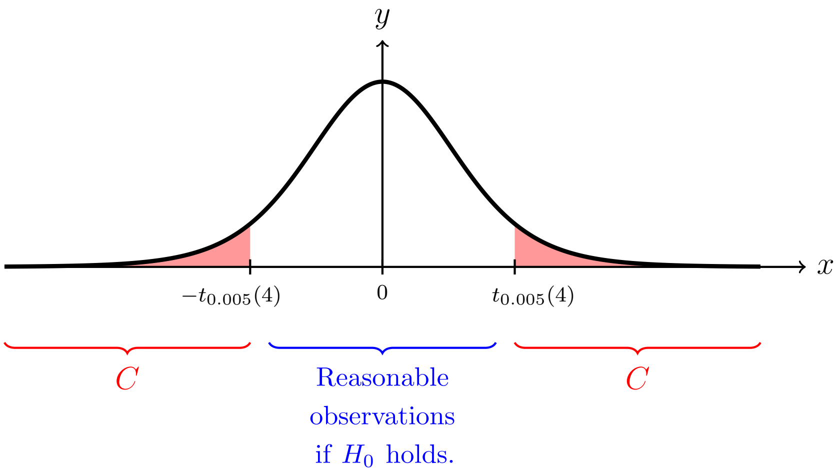
$$\widehat{X}(n) = \frac{2}{3}X(n-1) - \frac{1}{3}X(n-2).$$

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 $P(T \in C \mid H_0) = 0.01$ 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

$$H_0 : \beta_2 = 0,$$

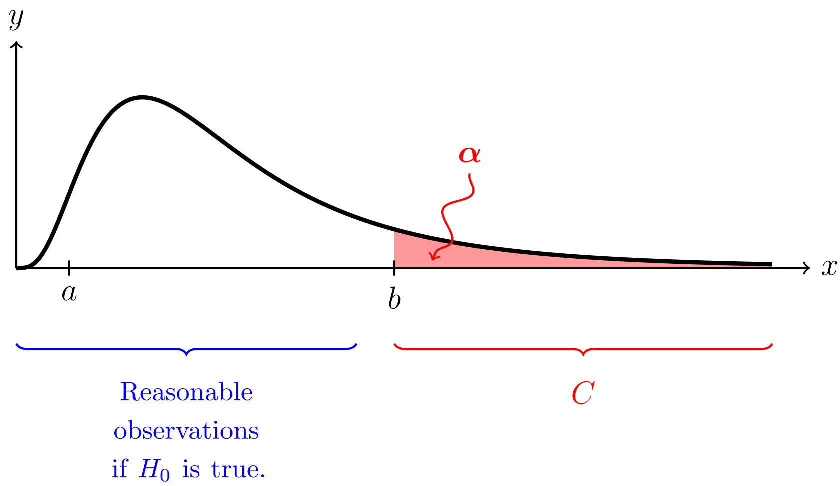
and

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

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

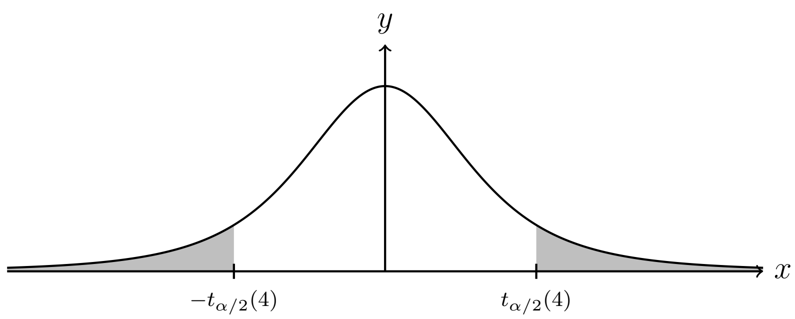
$$T = \frac{\widehat{\beta}_1 - \beta_1}{S\sqrt{h_{11}}} \sim t(4).$$

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

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

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):

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

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 $f(t) = \sum_{k=1}^{\infty} \frac{t^k}{k}$. 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

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

so the estimator is *not* unbiased.

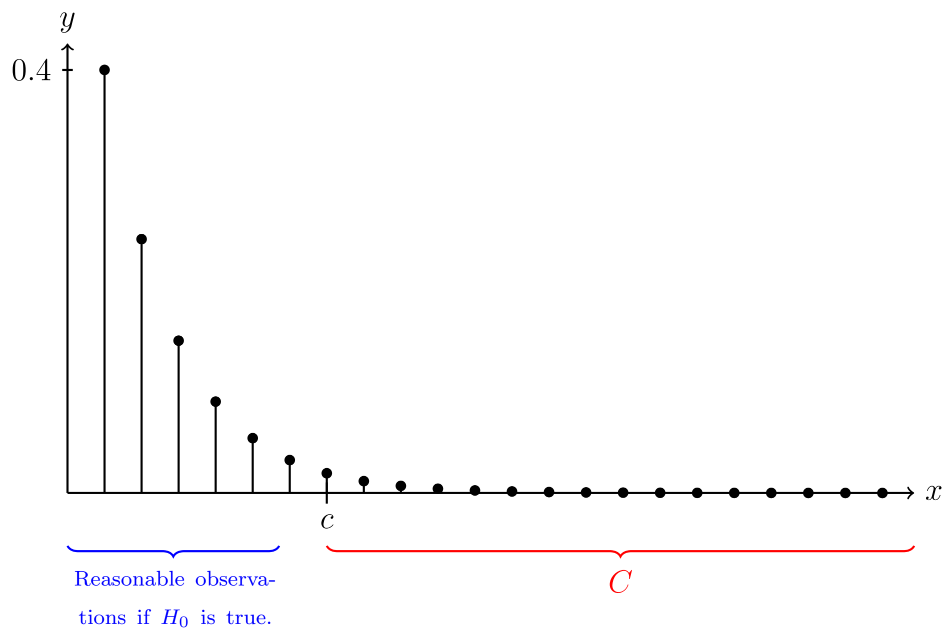
(b) Let X be the number of trials it takes for someone to finally get bitten. We concluded above that $X \sim \text{Ffg}(p)$, where p is the unknown probability of a bite. We want to test

$$H_0 : p = 0.4$$

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

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

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:

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

Answer: (a) $\widehat{P} = \frac{1}{x}$; 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: $\boldsymbol{Y}^T A\boldsymbol{Y} \sim \chi^2(l)$.