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

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

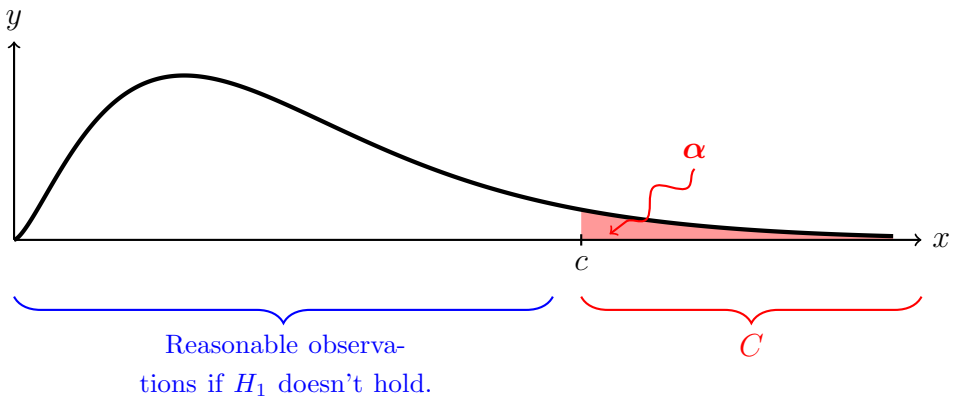
and

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

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)

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

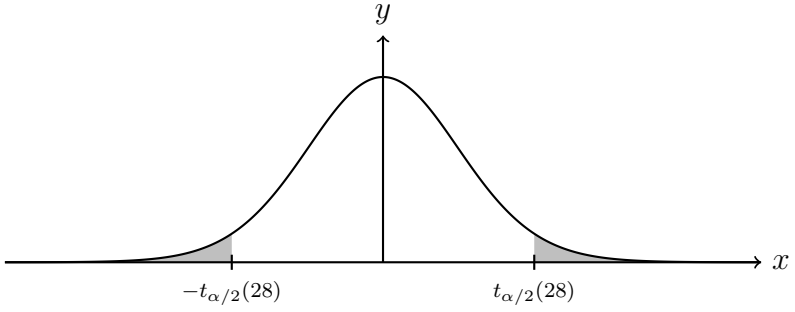
and

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

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.

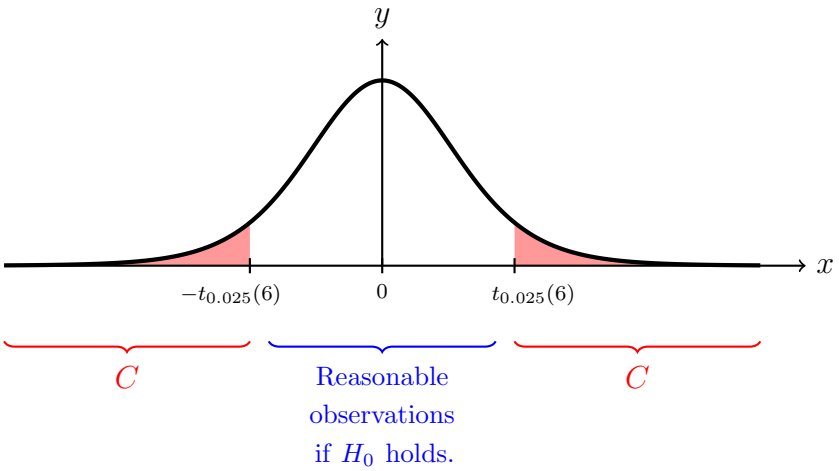
Answer: See above.

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.

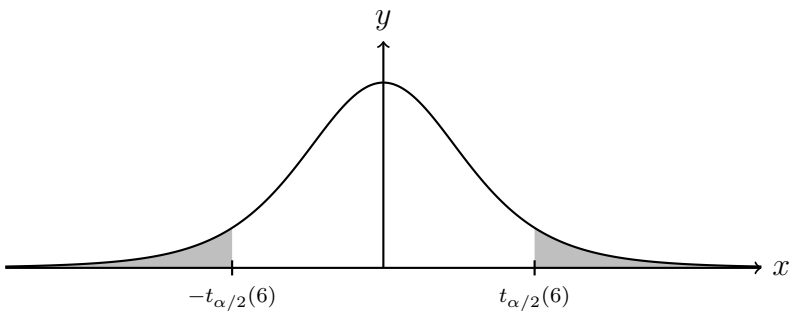
(c) To find a confidence interval for the expectation at $x = 0.5$, we let $u = (1\ 0.5\ 0.5^2)^T$. Let $\widehat{\mu_0} = \boldsymbol{u}^T\widehat{\boldsymbol{\beta}}$ be the estimate for μ at $x = 0.5$. Then (by Gosset's theorem)

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

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

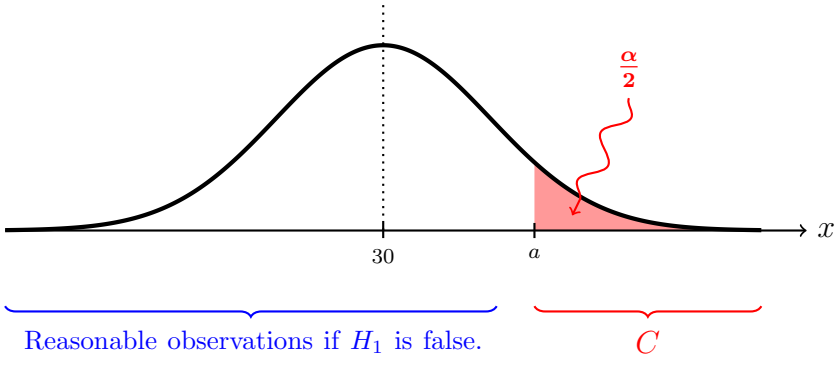
$$\overline{X} \sim N(30, 1.2^2/25) = N(30, 0.0576).$$

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.