Tufts University
Department of Mathematics
Spring 2022

MA 166: Statistics

Solutions to Homework 2 (v1.2) [1]

Assigned Monday 31 January 2022
Due Monday 7 February 2022 at 11:59 pm EDT.

Be sure to read the footnotes! Some of them are important.

1. **Suppose that you have a priori knowledge that the continuous random variable X is normally distributed. You make n experimental measurements of X, and you find that fn of the measurements are (miraculously) exactly equal to $+1$, and the other $(1-f)n$ measurements are exactly equal to -1. The fraction f is greater than $1/2$, so you think that X has a positive mean, but you are really not sure so you decide to do interval estimation to see with what confidence you can make that claim. You may assume that n is large enough that the Central Limit Theorem applies to a good approximation.**

   (a) **Use maximum likelihood estimators to find estimates of the mean, μ_e, and the standard deviation, σ_e.**

   The estimated mean is the sample mean,

   $$\mu_e = \frac{1}{n}\left[fn\,(+1) + (1-f)n\,(-1)\right],$$

   or

   $$\boxed{\mu_e = 2\left(f - 1/2\right),}$$

   which is greater than zero, since we are given that $f > 1/2$. The estimated standard deviation is

   $$\sigma_e = \sqrt{\frac{1}{n}\left[fn\left(+1-(2f-1)\right)^2 + n\left(-1-(2f-1)\right)^2\right]},$$

   or

   $$\boxed{\sigma_e = 2\sqrt{f(1-f)}.}$$

   (b) **We are going to use μ_e and σ_e as the basis for interval estimation in the remainder of this problem [2]. Suppose that you find you are able**

[2] As we shall see later, σ_e is not the optimum value to use for interval estimation, and the sample standard deviation for the normal distribution is usually defined in a different way. We have not yet covered that material however, so for the purposes of this problem you may proceed as instructed.

to conclude that the mean of X is positive with confidence probability $100(1-\alpha)\%$. Find an expression for n in terms of f and α [3]

We expect the quantity

$$Z = \frac{\frac{1}{n}\sum_{j=1}^{n} x_j - \mu}{\frac{\sigma_e}{\sqrt{n}}} = \frac{\mu_e - \mu}{\sigma_e}\sqrt{n}$$

to be distributed like a standard normal, and so we demand

$$\begin{aligned} 1-\alpha =& \mathrm{Prob}(\mu > 0) = \mathrm{Prob}\left(Z < \frac{\mu_e}{\sigma_e}\sqrt{n}\right) \\ =& \int_{-\infty}^{\frac{\mu_e}{\sigma_e}\sqrt{n}} dz\, \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right), \end{aligned}$$

so that

$$\alpha = \int_{\frac{\mu_e}{\sigma_e}\sqrt{n}}^{+\infty} dz\, \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right),$$

From this we may conclude

$$\begin{aligned} z_\alpha =& \frac{\mu_e}{\sigma_e}\sqrt{n} \\ =& \frac{2\left(f - {}^1\!/_2\right)}{2\sqrt{f(1-f)}}\sqrt{n}, \end{aligned}$$

or

$$\boxed{n = \frac{f(1-f)}{\left(f - {}^1\!/_2\right)^2}\, z_\alpha^2}$$

(c) **To make the above concrete, suppose that $f = 0.51$. How large does n need to be to achieve 95% confidence that the mean of X is positive?**

If we want $1-\alpha = 0.95$, then $\alpha = 0.05$, and from the table in Appendix A.1 (or your favorite mathematical software program), it is seen that $z_{0.05} = 1.64485\ldots$ The required n to achieve this confidence is then

$$n = \frac{(0.51)(1-0.51)}{(0.51-0.50)^2}(1.64485\ldots)^2,$$

or, rounding up to the next integer,

$$\boxed{n \approx 6762.}$$

It is seen that, even with a record of 51% measurements yielding $+1$ and 49% yielding -1, you could not conclude $E(X) > 0$ with 95% confidence unless you had taken at least 6762 samples.

[3] Feel free to use the z_α notation described in the text on page 298.

2. **Suppose that you have a priori knowledge that the continuous random variable $\boldsymbol{Y}$ has the following two-parameter probability density function:**

$$\boldsymbol{f_Y(y) = \begin{cases} \frac{1}{\mu-a}\exp\left(-\frac{y-a}{\mu-a}\right) & \text{for } y \geq a \\ 0 & \text{otherwise,} \end{cases}}$$

where it may be assumed that $\boldsymbol{a \geq 0}$ and $\boldsymbol{\mu > a}$.

(a) **Verify that the density function is normalized, and fine the theoretical mean and standard deviation of $\boldsymbol{Y}$ in terms of the parameters $\boldsymbol{a}$ and $\boldsymbol{\mu}$.**
Normalization is verified by the calculation

$$E(1) = \int_a^\infty dy \; \frac{1}{\mu - a} \exp\left(-\frac{y-a}{\mu-a}\right) = \int_0^\infty du \; \exp(-u) = 1,$$

where we made the u-substitution $u = \frac{y-a}{\mu-a}$. So

$$\boxed{E(1) = 1,}$$

demonstrating normalization.
The mean of Y is then

$$\begin{aligned} E(Y) &= \int_a^\infty dy \; \frac{1}{\mu-a} \exp\left(-\frac{y-a}{\mu-a}\right) y \\ &= \int_a^\infty du \; \exp(-u) \left[a + (\mu - a)u\right] = a + (\mu - a), \end{aligned}$$

or

$$\boxed{E(Y) = \mu.}$$

The mean of Y^2 is then

$$\begin{aligned} E(Y^2) &= \int_a^\infty dy \; \frac{1}{\mu-a} \exp\left(-\frac{y-a}{\mu-a}\right) y^2 \\ &= \int_a^\infty du \; \exp(-u) \left[a^2 + 2a(\mu-a)u + (\mu-a)^2u^2\right] = a^2 + 2a(\mu-a) + 2(\mu-a)^2 \end{aligned}$$

or

$$E(Y^2) = a^2 - 2a\mu + 2\mu^2,$$

and from this we can calculate $\sigma = \sqrt{E(Y^2) - [E(Y)]^2} = \sqrt{(\mu-a)^2}$, and taking the positive root,

$$\boxed{\sigma = \mu - a.}$$

(b) **Find maximum likelihood estimators, $\boldsymbol{\hat{a}_{\text{mle}}}$ and $\boldsymbol{\hat{\mu}_{\text{mle}}}$, for the parameters $\boldsymbol{a}$ and $\boldsymbol{\mu}$. Justify your reasoning carefully, particularly for the calculation of $\boldsymbol{\hat{a}_{\text{mle}}}$.**

The likelihood is given by

$$L(a, \mu; \vec{y}) = \prod_{j=1}^{n} f_Y(y_j).$$

This is clearly equal to zero unless $a \leq \min_j y_j$, so suppose this is the case. Then we have

$$L(a, \mu; \vec{y}) = \prod_{j=1}^{n} \frac{1}{\mu - a} \exp\left(-\frac{y_j - a}{\mu - a}\right) = \left(\frac{1}{\mu - a}\right)^n \exp\left(-\frac{\sum_{j=1}^{n} y_j - na}{\mu - a}\right)$$

which is positive, and the log likelihood is

$$\ln L(a, \mu; \vec{y}) = -n \ln(\mu - a) - \frac{1}{\mu - a}\left(\sum_{j=1}^{n} y_j - na\right).$$

Maximizing the log likelihood with respect to μ yields

$$0 = \frac{\partial}{\partial \mu} \ln L(a, \mu; \vec{y}) = -\frac{n}{\mu - a} + \frac{1}{(\mu - a)^2}\left(\sum_{j=1}^{n} y_j - na\right),$$

which may be solved to yield

$$\boxed{\mu_e = \frac{1}{n}\sum_{j=1}^{n} y_j,}$$

which conveniently does not depend on a. We may now consider

$$\begin{aligned}
\ln L(a, \mu_e; \vec{y}) &= -\, n \ln(\mu_e - a) - \frac{1}{\mu_e - a}\left(\sum_{j=1}^{n} y_j - na\right) \\
&= -\, n \ln(\mu_e - a) - \frac{1}{\mu_e - a}\left(n\mu_e - na\right)
\end{aligned}$$

which simplifies to

$$\ln L(a, \mu_e; \vec{y}) = -n - n \ln(\mu_e - a)$$

which is clearly an increasing function of a. Hence likelihood will be maximized when a is set to its minimum allowed value,

$$\boxed{a_e = \min_j y_j.}$$

This means that the estimated standard deviation is $\sigma_e = \mu_e - a_e$, or

$$\boxed{\sigma_e = \frac{1}{n}\sum_{j=1}^{n} y_j - \min_j y_j.}$$

(c) **Now suppose that you take n samples of this data, $y_1, \ldots, y_n$, where n is large enough that the Central Limit Theorem applies to a good approximation. (You may assume that the sample mean that you find is positive.) Find the $100(1-\alpha)\%$ confidence interval for $\mu = E(Y)$ centered at the sample mean. You may leave your answer in terms of $\hat{a}_{\text{mle}}$ and $\hat{\mu}_{\text{mle}}$ worked out in part (b).**

We may suppose that

$$Z = \frac{\mu_e - \mu}{\sigma_e/\sqrt{n}}$$

is distributed as a standard normal, where

$$\begin{aligned}
\mu_e &= \hat{\mu}_{\text{mle}}(\vec{y}) = \frac{1}{n}\sum_{j=1}^{n} y_j, \\
a_e &= \hat{a}_{\text{mle}}(\vec{y}) = \min_j y_j, \\
\sigma_e &= \mu_e - a_e = \frac{1}{n}\sum_{j=1}^{n} y_j - \min_j y_j.
\end{aligned}$$

Note that we use μ_e (and a_e) to compute σ_e, exactly as is done for the binomial parameter in Eq. (5.3.2) of the text. Hence we may write

$$z_{1-\alpha/2} \le Z \le z_{\alpha/2} \qquad \text{with confidence } 100(1-\alpha)\%,$$

or

$$z_{1-\alpha/2} \le \frac{\mu_e - \mu}{\sigma_e/\sqrt{n}} \le z_{\alpha/2} \qquad \text{with confidence } 100(1-\alpha)\%,$$

from which it is straightforward to confirm that the desired confidence interval is

$$\boxed{\mu \in \left[\mu_e - z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}},\ \mu_e + z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}}\right] \qquad \text{with confidence } 100(1-\alpha)\%.}$$

3. **Larsen & Marx, Problem 5.3.6, page 306**

   In this problem, we assume that Y is distributed normally with standard deviation σ.

   (a) **What is the confidence associated with the interval**

$$\left(\overline{y} - 1.64\frac{\sigma}{\sqrt{n}},\ \overline{y} + 2.33\frac{\sigma}{\sqrt{n}}\right)?$$

We have

$$\overline{y} - 1.64\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{y} + 2.33\frac{\sigma}{\sqrt{n}},$$

or

$$-1.64\frac{\sigma}{\sqrt{n}} \le \mu - \overline{y} \le +2.33\frac{\sigma}{\sqrt{n}},$$

or

$$-2.33\frac{\sigma}{\sqrt{n}} \leq \overline{y} - \mu \leq +1.64\frac{\sigma}{\sqrt{n}},$$

or

$$-2.33 \leq \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +1.64,$$

or

$$-2.33 \leq Z \leq +1.64,$$

where

$$Z := \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

By the Central Limit Theorem, we expect Z to be distributed as a standard normal. Consulting Appendix A.1, the above may be written

$$z_{0.9495} \leq Z \leq z_{0.0099},$$

and so the confidence associated with the interval is $0.9495 - 0.0099 = 0.9396$

$$\boxed{\text{Confidence} = 93.96\%}$$

(b) **What is the confidence associated with the interval**

$$\left(-\boldsymbol{\infty},\ \overline{\boldsymbol{y}} + \mathbf{2.58}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}}\right)\textbf{?}$$

We have

$$-\infty \leq \mu \leq \overline{y} + 2.58\frac{\sigma}{\sqrt{n}},$$

or

$$-\infty \leq \mu - \overline{y} \leq +2.58\frac{\sigma}{\sqrt{n}},$$

or

$$-1.64\frac{\sigma}{\sqrt{n}} \leq \overline{y} - \mu \leq +\infty,$$

or

$$-1.64 \leq \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq +\infty,$$

or

$$-1.64 \leq Z \leq +\infty,$$

where

$$Z := \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

By the Central Limit Theorem, we expect Z to be distributed as a standard normal. Consulting Appendix A.1, the above may be written

$$z_{1.0000} \leq Z \leq z_{0.0505},$$

and so the confidence associated with the interval is $1.0000 - 0.0505 = 0.9495$

$$\boxed{\text{Confidence} = 94.95\%}$$

(c) **What is the confidence associated with the interval**

$$\left(\overline{\boldsymbol{y}} - \mathbf{1.64}\frac{\boldsymbol{\sigma}}{\sqrt{\boldsymbol{n}}},\ \overline{\boldsymbol{y}}\right)?$$

We have

$$\overline{y} - 1.64\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{y},$$

or

$$-1.64\frac{\sigma}{\sqrt{n}} \le \mu - \overline{y} \le 0,$$

or

$$0 \le \overline{y} - \mu \le +1.64\frac{\sigma}{\sqrt{n}}$$

or

$$0 \le \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}} \le +1.64,$$

or

$$0 \le Z \le +1.64$$

where

$$Z := \frac{\overline{y} - \mu}{\frac{\sigma}{\sqrt{n}}}.$$

By the Central Limit Theorem, we expect Z to be distributed as a standard normal. Consulting Appendix A.1, the above may be written

$$z_{0.9495} \le Z \le z_{0.5000},$$

and so the confidence associated with the interval is $0.9495 - 0.5000 = 0.4495$

$$\boxed{\text{Confidence} = 44.95\%}$$

4. **Larsen & Marx, Problem 5.3.15, page 307**

   Suppose a coin is to be tossed $\boldsymbol{n}$ times for the purpose of estimating $\boldsymbol{p}$, where $\boldsymbol{p = P}$(heads). How large must $\boldsymbol{n}$ be to guarantee that the length of the 99% confidence interval for $\boldsymbol{p}$ will be less than 0.02?

   For $100(1-\alpha)\%$ to be equal to 99% confidence, we see that $\alpha = 0.01$. The standard deviation is

$$\sigma_e = \sqrt{p_e(1-p_e)},$$

   and we have

$$-z_{\alpha/2} \le \frac{p_e - p}{\sigma_e/\sqrt{n}} \le +z_{\alpha/2}.$$

   From this, it is straightforward to derive the confidence interval

$$p \in \left[p_e - z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}},\ p_e + z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}}\right],$$

and we can see that the width of the confidence interval is

$$\Delta p = \left(p_e + z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}}\right) - \left(p_e - z_{\alpha/2}\frac{\sigma_e}{\sqrt{n}}\right) = \frac{2z_{\alpha/2}\sigma_e}{\sqrt{n}}.$$

Solving for n, we find

$$n = \frac{4\sigma_e^2 z_{\alpha/2}^2}{(\Delta p)^2} = \frac{4p_e(1-p_e)z_{\alpha/2}^2}{(\Delta p)^2}.$$

This result depends on p_e, but we want it to work for all possible values of p_e. We know that the largest value that $4p_e(1-p_e)$ could possibly attain for $p_e \in [0,1]$ is one, which is achieved for $p_e = 1/2$. Hence the lowest value of n that is guaranteed to work for all p_e is

$$\boxed{n = \left(\frac{z_{\alpha/2}}{\Delta p}\right)^2.}$$

For $\alpha = 0.01$ and $\Delta p = 0.02$, we have $z_{0.01/2} = z_{0.005} = 2.575829\ldots$, so, rounding up to the nearest integer, if we choose

$$\boxed{n = \left(\frac{2.575829\ldots}{0.02}\right)^2 = 16588,}$$

we should always attain the desired interval width with the desired confidence.