GroupA_HM2

I.El gataa, M.Munini, E.Ruoppolo, M.Tallone

2022-10-20

Contents

FSDS - Chapter 3

Ex 3.12

Simulate random sampling from a normal population distribution with several n values to illustrate the law of large numbers

Solution

Add comments to the solution.

Ex 3.18

Sunshine City, which attracts primarily retired people, has 90,000 residents with a mean age of 72 years and a standard deviation of 12 years. The age distribution is skewed to the left. A random sample of 100 residents of Sunshine City has $\bar{y} = 70$ *and* $s = 11$.

- *Describe the center and spread of the (i) population distribution, (ii) sample data distribution. What shape does the sample data distribution probably have? Why?*
- *Find the center and spread of the sampling distribution of* $\overline{\mathrm{Y}}$ *for* $n = 100$*. What shape does it have and what does it describe?*
- *Explain why it would not be unusual to sample a person of age 60 in Sunshine City, but it would be highly unusual for the sample mean to be 60, for a random sample of 100 residents.*
- *Describe the sampling distribution of* $\overline{\mathrm{Y}}$ *: (i) for a random sample of size* $n = 1$; *(ii) if you sample all 90,000 residents.*

Solution

- The problem introduces the distribution of the entire population of Sunshine City, made by $N = 90,000$ residents. We can imagine it as a given discrete probability distribution from which we extract random variables by sampling it. With this procedure indeed we are randomly taking some elements from this distribution. As the number n of the values increase we expect that their distribution starts to be similar to the one from which they are sampled. So, considering $n = 100$ a sufficient sampling value, considering a completely random sampling procedure, the sample should be distributed as the population, and its centre and spread should be similar to the *real* values, that are in this case $\mu = 72$ and $\sigma = 12$, in fact they are 70 and 11. Of course because of the randomness of the sampling procedure the values of the centre and the spread of the finite sample could not be the same of the true values, but as n approaches to N so will do the values of centre and spread.
- We are now considering the sampling distribution of $\bar{\text{Y}}$. So we now have some different random samples each with $n = 100$ elements. The central limit theorem states that the distribution of the sample means, and in general of any sample statistics, will be approximately normal, no matter the shape of the population distribution. Clearly the samples for the applicability of the theorem it is required that the samples are all randomly taken from the population, that all the elements of each sample are independent and that the samples are large enough, and in our case $n = 100$ satisfies this condition. Since we suppose the other two condition to be true, the sample means will be approximately bell shaped, with expected value $\mu_{\bar{\text{Y}}} = \mu = 72$ years and standard deviation $\sigma_{\bar{\text{Y}}} = \sigma/\sqrt{n} = 1.2$ years. It says us that picking different random samples and each time computing the mean we expect that on average the mean will be 72 years, also these means will be normal distributed.
- It would not be strange to find a 60 years old person in Sunshine city. In fact the population distribution is left skewed, it means it has a long non-null "tail" on the left in which such an age, also considering the spread $\sigma = 12$ years that express the variability from the centre value, seems not strange. Anyway for a random sample of n finding a sample mean of 60 years would be really unusual for the reasons explained below. If the conditions of the CLT are all satisfied we expect a sample mean equal to the population mean and with a low variability $\sigma_{\bar{\text{Y}}} = 1.2$ years.
- In the first case we select every time a single value from the population. The mean is its self value, so practically we are randomly picking values from the population, they will be distributed approximately as the population itself. In the second case sampling the entire population means that each time the sample will always have mean value equal to the population one, so 72 years. Hence the sample distribution of means will not be a distribution at all since it will only have the same repeated value.

Ex 3.28

A survey is planned to estimate the population proportion π supporting more government action to address global warming. For a simple random sample, if π may be near 0.50, how large should n be so that the standard error of the sample proportion is 0.04?

Solution

Add comments to the solution.

FSDS - Chapter 4

Ex 4.2

Here the text of the first exercise.

Solution

Add comments to the solution.

Ex 4.4

For the `Students` *data file (Exercise 1.2 in Chapter 1) and corresponding population, find the ML estimate of the population proportion believing in life after death. Construct a Wald 95% confidence interval, using its* *formula (4.8). Interpret.*

Solution

Add comments to the solution.

Ex 4.38

For independent observations $y_1, \dots, y_n$ having the geometric distribution (2.1):

- *Find a sufficient statistic for π;*
- *Derive the ML estimator for π.*

Solution Let's first remind the geometric distribution:

$$f(y; \pi) = (1-\pi)^{y-1}\pi$$

- To find a sufficient statistic for the independent set $y_1, \dots, y_n$ from a geometric distribution we first need to write the likelihood function $L(\pi)$ that is:

$$L(\pi) = \prod_{i=1}^{n} (1-\pi)^{y_i - 1}\pi = \pi^n \prod_{i=1}^{n} (1-\pi)^{y_i-1} = \left(\frac{\pi}{1-\pi}\right)^n \prod_{i=1}^{n} (1-\pi)^{y_i} = \left(\frac{\pi}{1-\pi}\right)^n (1-\pi)^{\sum_i^n y_i}$$

  Hence the log likelihood $\ell(\pi)$ is:

$$\ell(\pi) = \log L(\pi) = n[\log \pi - \log(1-\pi)] + \log(1-\pi) \sum_i^n y_i$$

  Since we can write it as a factorized function of type $g[T(\mathbf{y}); \pi]$ where $T(\mathbf{y}) = \sum_i^n y_i$ and $\ell(\pi)$ depends on the data $\mathbf{y}$ only through this function that comprehends the entire data, and not single observations, we can say that $T(\mathbf{y}) = \sum_i^n y_i$ is a sufficient statistic.

- To find an ML estimator for π we have to find a value of π that maximize $\ell(\pi)$. So we can find the value of π for which the derivative of ℓ is null:

$$\frac{\mathrm{d}}{\mathrm{d}\pi} \ell(\pi) = n \left(\frac{1}{\pi} + \frac{1}{1-\pi}\right) - \frac{1}{1-\pi} \sum_i^n y_i = 0$$

  Hence:

$$\frac{1}{n} \sum_i^n y_i = \frac{1-\pi+\pi}{\pi} \implies \frac{1}{\pi} = \frac{1}{n} \sum_i^n y_i$$

  Thus we find the ML estimator for π:

$$\hat{\pi} = \frac{n}{\sum_i^n y_i}$$

Ex 4.44

Refer to the previous two exercises. Consider the selling prices (in thousands of dollars) in the `Houses` *data file mentioned in Exercise 4.31.*

- *Fit the normal distribution to the data by finding the ML estimates of μ and σ for that distribution.*
- *Fit the log-normal distribution to the data by finding the ML estimates of its parameters.*
- *Find and compare the ML estimates of the mean and standard deviation of selling price for the two distributions.*
- *Superimpose the fitted normal and log-normal distributions on a histogram of the data. Which distribution seems to be more appropriate for summarizing the selling prices?*

Solution

Add comments to the solution.

Ex 4.48

For a simple random sample of n *subjects, explain why it is about 95% likely that the sample proportion has error no more than* $1/\sqrt{n}$ *in estimating the population proportion. (Hint: To show this "* $1/\sqrt{n}$ *rule", find two standard errors when* $\pi = 0.50$ *, and explain how this compares to two standard errors at other values of* π *.) Using this result, show that* $n = 1/M^2$ *is a safe sample size for estimating a proportion to within* M *with 95% confidence.*

Solution

Add comments to the solution.

FSDS - Chapter 5

Ex 5.2

When a government does not have enough money to pay for the services that it provides, it can raise taxes or it can reduce services. When the Florida Poll asked a random sample of 1200 Floridians which they preferred, 52% (624 of the 1200) chose raise taxes and 48% chose reduce services. Let π *denote the population proportion of Floridians who would choose raising taxes.Analyze whether this is a minority of the population* $(\pi < 0.50)$ *or a majority* $(\pi > 0.50)$ *by testing* $H_0 : \pi = 0.50$ *against* $H_\alpha : \pi \neq 0.50$ *. Interpret the P-value. Is it appropriate to "accept"* H_0 *? Why or why not?*

Solution

Add comments to the solution.

Ex 5.12

The example in Section 3.1.4 described an experiment to estimate the mean sales with a proposed menu for a new restaurant. In a revised experiment to compare two menus, on Tuesday of the opening week the owner gives customers menu A and on Wednesday she gives them menu B. The bills average $22.30 for the 43 customers on Tuesday (s = 6.88) and $25.91 for the 50 customers on Wednesday (s = 8.01). Under the strong assumption that her customers each night are comparable to a random sample from the conceptual population of potential customers, show how to compare the mean sales for the two menus based on (a) the P-value of a significance test, (b) a 95% confidence interval. Which is more informative, and why? (When used in an experiment to compare two treatments to determine which works better, a two-sample test is often called an A/B test.).

Solution

a) To compare the mean sales for the two menus with a significance test, given the assumption made by the problem we can use the either the standard t-test, with the pooled standard deviation. First of all we estimate a pooled value s for the standard deviation:

$$s = \sqrt{\frac{(n_1 - 1)s_1^2 + (n2_1)s_2^2}{n_1 + n_2 - 2}}$$

This is possible since s_1 and s_2 do not result greatly different. So in this case the estimated standard error is:

$$se = s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

The test consider as null hypothesis $H_0 : \mu_1 = \mu_2$, so $\mu_1 - \mu_2 = 0$, otherwise $H_\alpha : \mu_1 \neq \mu_2$. Hence the test statistic is:

$$T = \frac{(\overline{Y}_1 - \overline{Y}_2) - 0}{SE}$$

and df is $df = n_1 + n_2 - 2$. So considering $(\overline{Y}_1, s_1, n_1) = (22.30, 6.88, 43)$ and $(\overline{Y}_2, s_2, n_2) = (25.91, 8.01, 50)$ we can compute:

```
y1 <- 22.30
s1 <- 6.88
n1 <- 43
y2 <- 25.91
s2 <- 8.01
n2 <- 50

degf = n1+n2-2

s = sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(degf))
se = s * sqrt(1/n1+1/n2)

t = (y1-y2)/se

se; t; degf
```

```
## [1] 1.561853
```

```
## [1] -2.311357
```

```
## [1] 91
```

```
pvalue = 2*pt(q= t, df = degf, lower.tail = T)
pvalue
```

```
## [1] 0.02307139
```

We can estimate the p-value as $p = 0.02$. This low value suggests that we should reject H_0 hence $\mu_1 \neq \mu_2$ and the mean sales between the two menus are different.

b) Now we use a 95% confidence interval to evaluate the difference between μ_1 and μ_2. We now repeat the procedure already used above. The confidence interval for $\mu_1 - \mu_2$ is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{\alpha/2,\, n_1+n_2-2}(se)$$

In our case $\alpha/2 = (1 - 0.95)/2 = 0.025$, $df = n_1 + n_2 - 2 = 91$ thus we can compute the quantities:

```
y1 <- 22.30
s1 <- 6.88
n1 <- 43
y2 <- 25.91
s2 <- 8.01
n2 <- 50
s = sqrt(((n1-1)*s1^2+(n2-1)*s2^2)/(degf))
se = s * sqrt(1/n1+1/n2)
m_dif=y1-y2
t = qt(0.025, 91, lower.tail = F)
dif = t*se
m_dif
```

```
## [1] -3.61
```

```
dif
```

```
## [1] 3.102429
```

The result is:

$$(\bar{y}_1 - \bar{y}_2) \pm t_{0.025,91}(se) = -3.61 \pm 3.1 \text{ or } (-6.71, -0.51)$$

The interval not including 0 can be interpreted as sign that the two means are not equal. Also the interval covering only negative values may mean that $\mu_2 > \mu_1$.

So, by these two different analysis we can state that the hypothesis test only states that the hypothesis of having $\mu_1 = \mu_2$ should be rejected because of the low $p-$value while the confidence interval analysis adds to the statement of incompatibility that the mean sale of menu b is higher than the one of menu a.

Ex 5.68

Explain why the confidence interval based on the Wald test of $H_0 : \theta = \theta_0$ *is symmetric around* $\hat{\theta}$ *(i.e., having center exactly equal to* $\hat{\theta}$ *. This is not true for the confidence intervals based on the likelihood-ratio and score tests.) Explain why such symmetry can be problematic when* θ *and* $\hat{\theta}$ *are near a boundary, using the example of a population proportion that is very close to 0 or 1 and a sample proportion that may well equal 0 or 1.*

Solution

Add comments to the solution.