Problem set 1 model solutions

Tuomas Markkula

1/14/2022

These are **suggested solutions** and we have tried to explain the solutions to you thoroughly. Your answers should be much more concise. It is possible to get high points even if it seems to you that your answers differ from the suggested solutions.

Exercise 1

(8p) Compute the following probabilities:

- $P(6 \leq Y \leq 12)$ when $Y \sim N(10, 64)$. Use a standard normal distribution. Demonstrate how you do this.
- $P(12 \leq Y)$ when $Y \sim N(14, 36)$. Use a computer for the calculations and report how you do this.
- $P(Y > 1.75)$ when Y follows Student's t-distribution with 15 degrees of freedom. What kind of a variable has this distribution? What happens when the parameter of the Student's t-distribution gets large?

Solution

a)

In this question we first normalize the $Y \sim N(10, 64)$ and then calculate the probability. Let us first normalize.

$$P(6 \leq Y \leq 12) = P\left(\frac{6-10}{\sqrt{64}} \leq \frac{Y-10}{\sqrt{64}} \leq \frac{12-10}{\sqrt{64}}\right) \tag{1}$$

$$= P\left(-\frac{1}{2} \leq \frac{Y-10}{8} \leq \frac{1}{4}\right) \tag{2}$$

$$= P\left(\frac{Y-10}{8} \leq \frac{1}{4}\right) - P\left(-\frac{1}{2} \leq \frac{Y-10}{8}\right) \tag{3}$$

Because $\frac{Y-10}{8} \sim N(0, 1)$ we can use the standard normal cumulative distribution function to calculate the probability.

$$\Phi(0.25) - \Phi(-0.5) = 0.29$$

Using R:

```
# pnorm() gives the cumulative distribution function of a normally distributed RV.
pnorm(1/4, mean = 0, sd = 1) - pnorm(-1/2, mean = 0, sd = 1)
```

```
[1] 0.2901688
```

Stata code:

```
di normal(1/4) - normal(-1/2)
```

.29016879

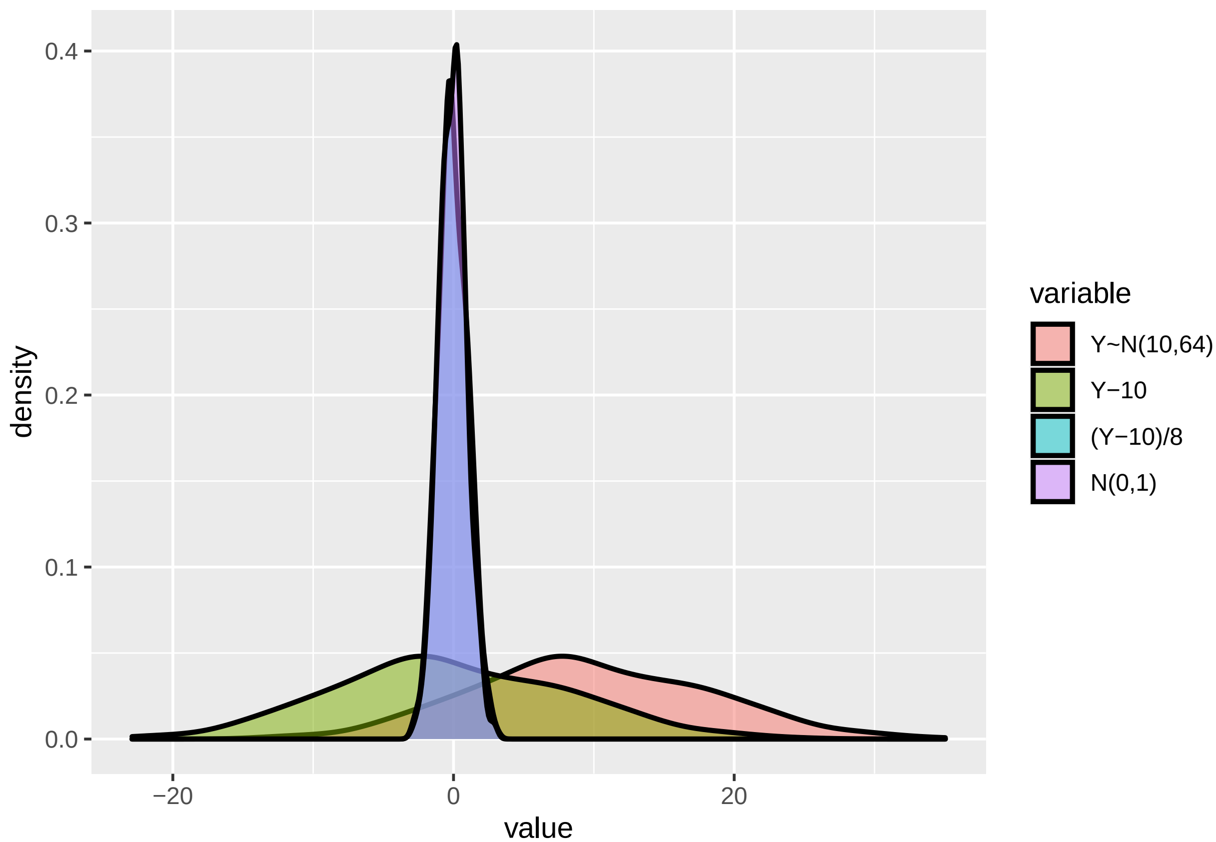
```
normal <- rnorm(500, mean = 10, sd = sqrt(64))

std_normal <- rnorm(500)

dt <-
  data.table("Y~N(10,64)" = normal,
             "Y-10" = normal - 10,
             "(Y-10)/8" = (normal - 10)/sqrt(64),
             "N(0,1)"= std_normal)

dt <- melt(dt)

ggplot(dt, aes(x = value, fill = variable)) +
  geom_density(alpha = 0.5, size = 0.9, linetype = 1)
```



b)

In this question you were asked to calculate the probability using a computer. Now we don't need to standardize the distribution if we do not want. We can just use computer to give us Y's distribution $N(14, 36)$ and use that directly. Alternatively we can proceed as in a). Note that $P(12 \leq Y) = P(Y \geq 12) = 1 - P(Y < 12) = 0.63$ (This is bad notation on our part, typically we always denote $P(Y \leq a)$ $P(Y > a)$ and not $P(a < Y)$). In R we can calculate the probability in the following way:

```r
# pnorm() gives the cumulative distribution function of a normally distributed RV.
# with parameters mean and standard deviation.
1 - pnorm(12, mean = 14, sd = 6)
```

```
[1] 0.6305587
```

Stata code (normalization is required in Stata, as there is no similar function as in R):

```
di 1 - normal((12-14)/6)
```

```
.63055866
```

c)

Now we need to use Student's t-distribution with 15 degrees of freedom. I just pick another R function that gives me the wanted distribution. Note the direction of the inequality and that $P(Y > 1.75) = 1 - P(1.75 \leq Y)$.

$$1 - P(1.75 \leq Y) = 0.05$$

```r
# pt() gives the cumulative distribution function of a Student's t-distributed RV.
# with parameter degrees of freedom
1 - pt(1.75, df = 15)
```

```
[1] 0.05027009
```

Stata code:

```
di 1 - t(15, 1.75)
```

```
.05027009
```

What kind of variable has this distribution? Take a sample of n observations from distribution $N(\mu, \sigma^2)$. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Then

$$\frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t\text{-distribution}_{n-1}.$$

When the parameter of the Student's t-distribution gets larger than 30 the Student t is well approximated by the standard normal distribution. When $n-1$ tends to infinity then the t-distribution tends to the standard normal distribution.

Exercise 2

(8p) What is the distribution of X when

- $Y_i \sim N(\mu, \sigma^2)$, Y_is are independent and $X = Y_1 + Y_2 + Y_3$.
- $Y_1, Y_2, Y_3 \sim N(0, 1)$, Y_is are independent and $X = Y_1^2 + Y_2^2 + Y_3^2$.
- $Y_1 \sim \chi^2_4$ and $Y_2 \sim \chi^2_8$, Y_is are independent and $X = \frac{Y_1/4}{Y_2/8}$. What happens when the parameter of the Y_2 tends to infinity?

Solution

a)

$X \sim N(\mu_1 + \mu_2 + \mu_3,\ \sigma_1^2 + \sigma_2^2 + \sigma_3^2)$

b)

$X \sim \chi^2_3$

c)

$X \sim F_{4,8}$.

What happens when the parameter of the Y_2, m, tends to infinity? Because $Y_2 = \sum_{i=1}^{m} Y_{i2}^2$, where m is the degrees of freedom of the Chi-squared distribution. Thus as the m tends to infinity the ratio Y_2/m tends to 1. Consequently

$$X = \frac{Y_1/4}{Y_2/m} \to Y_1/4 \text{ as } m \to \infty$$

Why the ratio Y_2/m tends to 1 when the m tends to ∞.

By the definition of Y_2

$$Y_2/m = \frac{1}{m}\sum_{i=1}^{m}(Y_{i2})^2 \tag{4}$$

Where $Y_{i2} \sim N(0, 1)$. When $m \to \infty$ then by the law of large numbers

$$Y_2/m = \frac{1}{m}\sum_{i=1}^{m}(Y_{i2})^2 \to E[Y_{i2}^2] \text{ in probability.} \tag{5}$$

From the simple formula of the variance

$$Var(Y) = E(Y^2) - (E(Y))^2 \implies E(Y^2) = Var(Y) + (E(Y))^2$$

Then

$$E[Y_{i2}^2] = \sigma^2 + \mu_{Y_2}^2 = \sigma^2 + 0 = 1$$

because $Y_{i2} \sim N(0, 1)$. Hence as m tends to infinity

$$Y_2/m \to 1 \text{ in probability.}$$

Exercise 3

(12p) In this study we analyze the results of a survey by The Finnish Business School Graduates (FBSG), a central organization for graduates and students in economics and and business administration, on summer job pay. The survey targeted student members of the FBSG who were at least on their second year of studies and had begun their studies after 2009. The survey was conducted as an electronic survey that was e-mailed to the eligible members of the FBSG in the autumn of 2021. The above table is a translation of the Table 3 in the original report of the survey results by Venäläinen (2021). The table presents mean gross monthly salary from a summer job by the number of credits. You can assume that the observations are iid.

Credits		under 120	120-180	181-220	over 220
Observations		(N: 282)	(N: 251)	(N: 180)	(N: 179)
Gross monthly	Mean	2 384	2 436	2 514	2 419
salary	(Sd)	(1 151)	(1 289)	(1 100)	(804)
	Median	2 100	2 100	2 300	2 300

- Is there evidence that students with over 220 credits earn less on average than those who have 181-220 credits? Use a p-value.
- Calculate a 95% confidence interval for the mean gross monthly salary for the students with 120-180 credits. What does the confidence interval imply?
- Compare means and medians in the table. What does this comparison tell you about the data?

Solution

a)

In this question we need to do a one sided t-test to test difference in the means between the two groups of students. Lets index students with over 220 credits with 1 and the other group with 2. Our null hypothesis is that students with 220 credits earn more or as much as students with 181-220 credits: $H_0 : \mu_1 - \mu_2 \geq d_0$ with $d_0 = 0$, where μ_i is the average salary of group i. The alternative hypothesis is that the group 1 earns less than the group 2 on average: $H_1 : \mu_1 - \mu_2 < d_0$.

The formula for the t-test is

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{SE(\overline{Y_1} - \overline{Y_2})}$$

where

$$SE(\overline{Y_1} - \overline{Y_2}) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Because the t-test static has a standard normal distribution under the null hypothesis when n_1 and n_2 are large enough, we can use the above formula to calculate the test static and standard normal distribution to get the p-values.

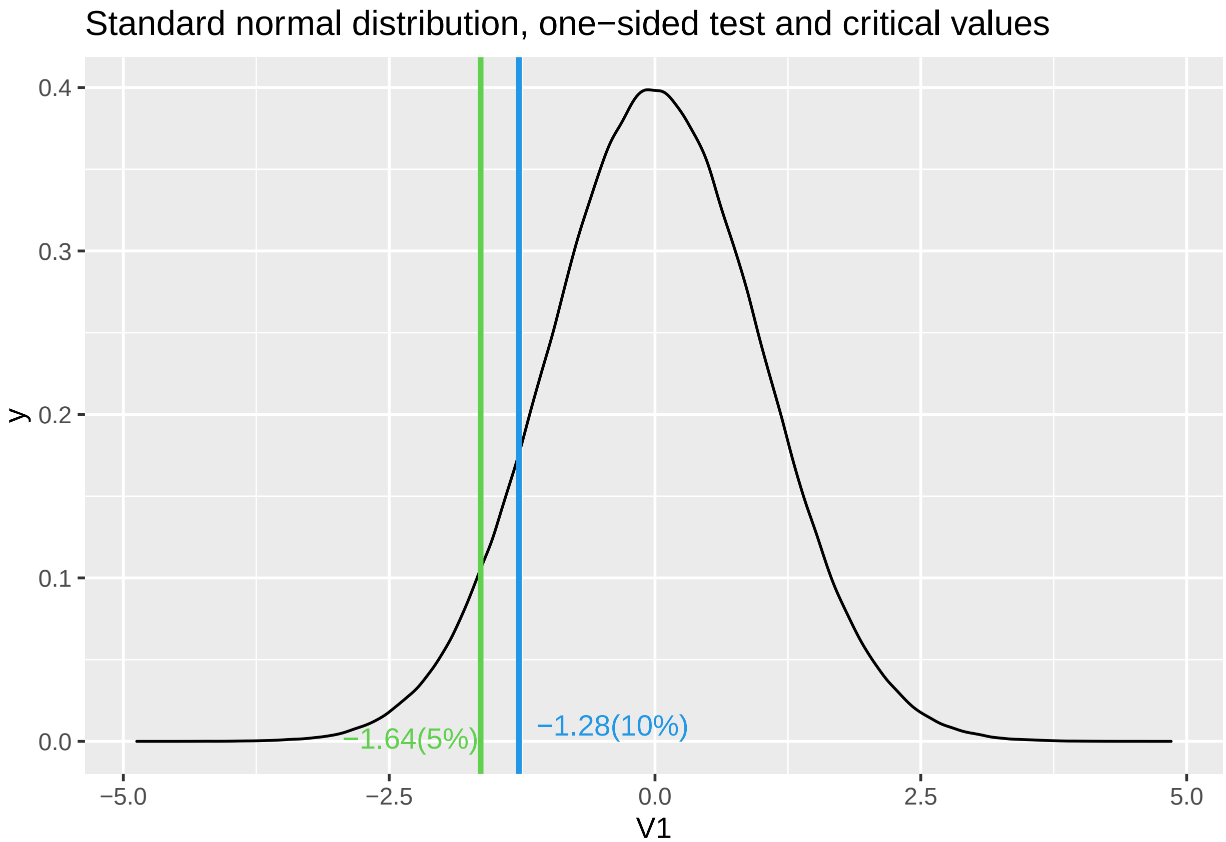
$$t = \frac{2419 - 2514}{\sqrt{\frac{804^2}{179} + \frac{1100^2}{180}}} = -0.935$$

The one sided critical value at 5% significance level for our one sided hypothesis is -1.64. As our t-test static is larger than the threshold value we are not able to reject the null hypothesis that the mean salary of the group 1 is equal or larger than the mean salary of the group 2. The p-value we get from $P(T \leq -0.935)$, $T \sim N(0,1)$, under the null is 0.17 which is larger than the 0.05, so we can't reject the null hypothesis. The same holds at the 10% significance level, where the critical value is -1.28.

When to use a one-sided t-test and when to use a two-sided t-test. Discussion on the topic here.

```
dt <- data.table(rnorm(1000000))

ggplot(dt, aes(x=V1)) +
  geom_density() +
  geom_vline(xintercept = -1.64, col = 3, size = 1) +
  annotate("text", x = -2.3, y = 0.002, col = 3, label = "-1.64(5%)")+
  geom_vline(xintercept = -1.28, col = 4, size = 1) +
  annotate("text", x = -0.4, y = 0.01, col = 4, label = "-1.28(10%)")+
  ggtitle("Standard normal distribution, one-sided test and critical values")
```



b)

As we are working with a relatively large sample we can use a normal approximation which leads to the following formula for the 95% confidence interval. Especially, the number 1.96 comes from the standard normal distribution and the fact that we want a 95% confidence interval.

$$\overline{Y} \pm 1.96 \times SE(\overline{Y})$$

where th $SE(\overline{Y}) = \hat{\sigma} = s_Y/\sqrt{n}$.

We get

$$\overline{Y} \pm 1.96SE(\overline{Y}) = 2436 \pm 1.96 \times 81.36 = [2277,\ 2595]$$

Because the confidence interval contains the means of all groups in the table, there is no evidence that the mean summer job salary for the students with 120-180 would be statistically significantly different from the students in the other credit brackets at the 5% significance level.

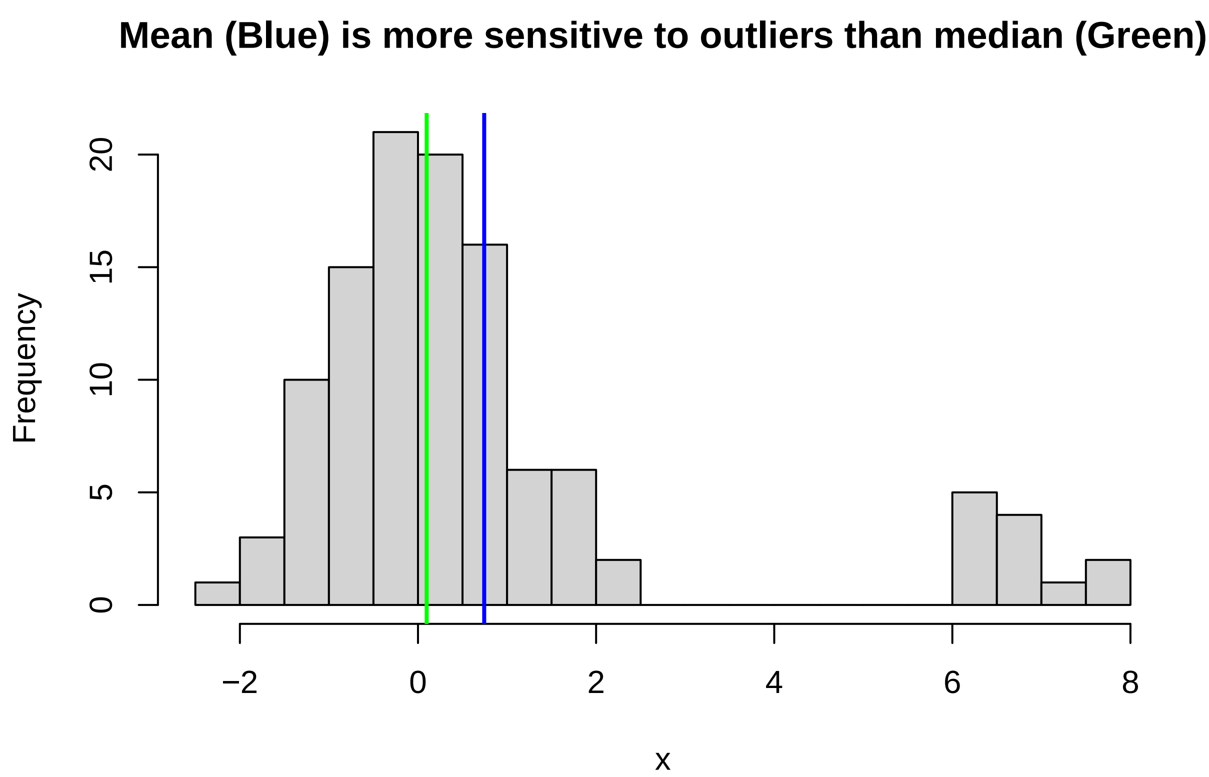
c)

As the median is smaller than the mean it seems that the data is skewed towards the right tail. Mean is more suspect to outliers in the tail, while the median captures the central tendency of the data more reliably. In this case no-one has salary lower than 0, while it is likely that some individuals have salaries substantially higher than the median.

```
x <- rnorm(100)
outliers <- runif(12, 6, 8)

x <- c(x, outliers)
```

```
hist(x, breaks = 25, main = "Mean (Blue) is more sensitive to outliers than median (Green)")
abline(v = mean(x), col = "blue", lwd = 2)
abline(v = median(x), col = "green", lwd = 2)
```



Exercise 4

(4p) What do you think about the following statement found when one asks Google "What confidence interval means?"

Solution

The statement is incorrect. A confidence interval does not give one a probability that a parameter falls between a pair of values. Rather, the interpretation of a CI is that in repeated sampling in 95% of the samples the true parameter value is within the confidence interval and in 5% of the samples outside the interval. It is wrong to say that the true parameter value is within the interval with 95% probability, as in each sample the parameter either is within the interval or not.

Exercise 5

(16p) The data on the body heights of ECON-C4100 participants yields the mean of $\overline{Y} = 177.04$ and standard deviation of $s_Y = 10.80$ with $n = 45$.

(a) Construct a 95% confidence interval for the mean body height in the population.

(b) What is the relationship between the 95% confidence level and a 5% significance level?

Solution

a) Similarly as in Exercise 4. we use

$$\overline{Y} \pm 1.96 \times SE(\overline{Y})$$

where $SE(\overline{Y}) = \hat{\sigma} = s_Y/\sqrt{n}$. Let us plug in the numbers

$$177.04 \pm 1.96 \times \frac{10.80}{\sqrt{45}} = [173.88, 180.20]$$

b)

By choosing a significance level 5% one obtains the 95% confidence interval. The 95% confidence interval contains all the values that cannot be rejected at the 5% significance level.

Exercise 5 continues c), d)

The sample average and standard deviation for the men are $\overline{X} = 183.00$ and $s_X = 7.52$. The sample average and standard deviation for the women are $\overline{Y} = 165.13$ and $s_Y = 4.67$. The number of men participants in the survey is 30 and the number of women participants is 15.

- c) Is there statistically significant evidence that the average body height of men is higher than the average body height of women?

A course TA calculated the sample averages and standard deviations for the first half ($N = 23$) and for the second half ($N = 22$) of the survey participants separately. The results yield $\overline{X} = 178.48$, $s_X = 10.38$ and $\overline{Y} = 175.55$ and $s_Y = 11.27$.

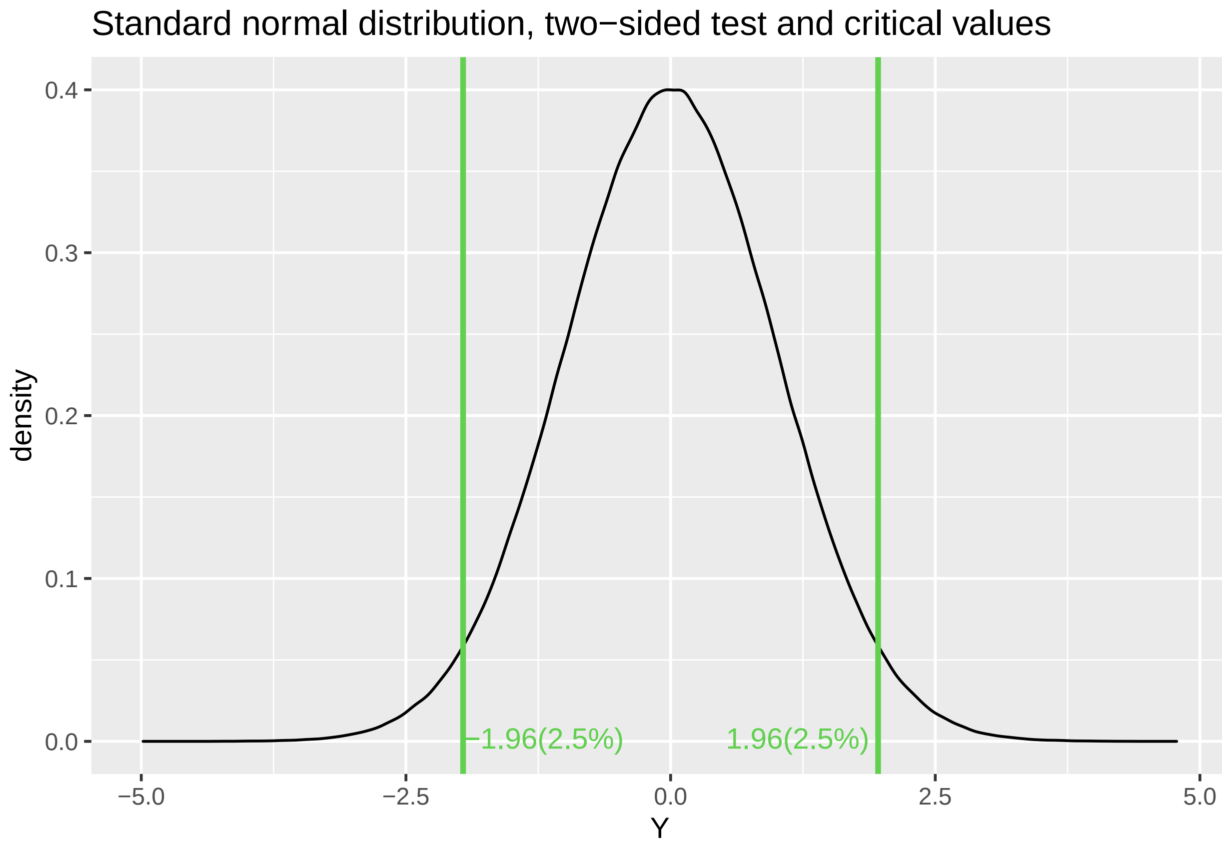
- d) Is there statistically significant evidence that the average body height of the first subgroup is higher than the average body height of the second group? Do you find your result plausible and why?

Solution

c) and d)

```
dt <- data.table(rnorm(1000000))

ggplot(dt, aes(x=V1)) +
  geom_density() +
  geom_vline(xintercept = -1.96, col = 3, size = 1) +
  geom_vline(xintercept =  1.96, col = 3, size = 1) +
  annotate("text", x = -1.2, y = 0.002, col = 3, label = "-1.96(2.5%)")+
  annotate("text", x =  1.2, y = 0.002, col = 3, label =  "1.96(2.5%)") +
  ggtitle("Standard normal distribution, two-sided test and critical values") +
  labs(x="Y", y = "density")
```



Note here that we could have formulated our hypothesis test as a one-sided test like in the Exercise 3. Let us consider a two-sided test here, as we have already discussed the one-sided alternative above. Both implemented correctly give full points. We have the following null hypothesis: $H_0 : \mu_1 - \mu_2 = d_0$ versus the alternative hypothesis $H_1 : \mu_1 - \mu_2 \neq d_0$ and $d_0 = 0$. There are two possible methods for the general solution:

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - d_0}{SE(\overline{Y}_1 - \overline{Y}_2)}$$

Where

$$SE(\overline{Y}_1 - \overline{Y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Thus, the test statistic is calculated as

$$t = \frac{(\overline{Y}_1 - \overline{Y}_2) - d_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

If the sample sizes are large enough, t follows a standard normal distribution. Otherwise it follows the Student's distribution with roughly

$$d.f. \approx (\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2/(\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_2^4}{n_2^2(n_2-1)})$$

degrees of freedom. In this exercise, you will get full points for using both distributions. The t-statistic formula must be correct though. At the 5% significance level, we will reject H_0 if $|t| > 1.96$ in the case of the standard normal distribution and if $|t| >\approx 2.00$.

In part c), we have $t \approx 9.78$ using a standard normal distribution. Thus, we will reject the null hypothesis that the difference of the means is zero.

In part d), we have $t \approx 0.91$. Thus, we cannot reject the null hypothesis that the difference of the group means is zero.

Exercise 6

(16p) Suppose you throw a coin

- ten times to obtain your sample. What does a sample being iid mean? Is your sample iid?
- ten times. Assume that the coin is fair. If head $Y_i = 1$ and if tails $Y_i = 0$. What is the probability that less than two out of the ten throws are heads ($P(\overline{Y} \leq 0.2)$)? Approximate using the Central Limit Theorem defined in Lecture 2B or in the book.
- hundred times. Assume that the coin is fair. What is the probability that the coin gives heads with a probability less ($\leq$) than 1/4 ($P(\overline{Y} \leq 1/4)$)? Approximate using the CLT.
- Why you can use the CLT? What happens to your approximation as the sample size increases and why?

Solution

a)

A sample being an iid means that the observations are identically distributed and independent. The former means that all of the observations have the same distribution. The latter means that the observations are not in any way related to each other. The sample drawn here is iid as we are throwing the same coin over and over again and hence the probability of getting heads is always the same, and thus the distribution is also the same. Observations are independent because any of the the previous throws has no effect on the following throw and the following throws have no effect on the previous throws.

b)

By the CLT the random variable

$$\frac{\overline{Y} - \mu_Y}{\sigma_{\overline{Y}}} = \frac{\overline{Y} - \mu_Y}{\sigma_Y/\sqrt{n}}$$

should be well approximated by the standard normal distribution when sample n gets large. Using this results let us calculate the probability $P(\overline{Y} < 0.25)$ when our sample n is 10.

Our random variable Y_i has Bernoulli distribution, which has expected value $E[Y_i] = \mu = p$ where p is the share of heads out of all throws, and variance $Var(Y_i) = p(1-p) = \sigma^2$. As the coin is fair $p = 0.5$. We have

$$E[Y_i] = \mu_Y = 0.5$$
$$Var(Y_i) = p(1-p) = 0.25$$

Using these we have

$$P(\overline{Y} \leq 0.2) = P\left(\frac{\overline{Y} - 0.5}{0.16} \leq \frac{0.2 - 0.5}{0.16}\right) = P\left(\frac{\overline{Y} - 0.5}{0.16} \leq \frac{-0.3}{0.16}\right)$$

Using the standard normal CDF we obtain

$$\Phi(-0.3/0.16) = 0.03$$

c)

The main difference from b) is that we have larger sample. We want to calculate

$$P(\overline{Y} \leq 0.25) = P\left(\frac{\overline{Y} - 0.5}{0.05} \leq \frac{0.25 - 0.5}{0.05}\right) = P\left(\frac{\overline{Y} - 0.5}{0.05} \leq \frac{-0.25}{0.05}\right)$$

Using the standard normal CDF we get

$$\Phi(-0.25/0.05) = 2.866516 \times 10^{-7} \approx 0$$

d)

We can use the CLT because the observations are iid and the variance is finite, meaning that $0 < \sigma_Y^2 < \infty$. The approximation gets better because the CLT says that as the sample size n tends to infinity the distribution of the standardized sample average becomes arbitrarily well approximated by a standard normal distribution. From a practical point of view, we always have a finite sample. Unfortunately, the CLT does not say when the sample is big enough so that the approximation is good in some sense. So most of the time we just hope that our sample is large enough.

Exercise 7

(18p) Consider two jointly distributed random variables Y and Q. You know the value of Y, while the value for Q is unknown. Denote a guess of the value of Q by $\tilde{Q} = E[Q|Y]$ that uses the information on Y. Finally the $V = Q - \tilde{Q}$ is the error of this guess. Remember to state what assumptions or properties you use and where.

- Show that $E[V] = 0$.
- Show that $E[VY] = 0$.
- $\hat{Q} = g(Y)$ is another guess of Q using Y, and $W = Q - \hat{Q}$ is its error. Show that $E[W^2] \geq E[V^2]$.

Hints: (a) Law of iterated expectations. (c) Use $h(Y) = g(Y) - E[Q|Y]$ such that $W = (Q - E[Q|Y]) - h(Y)$. Obtain the expression for the $E[W^2]$.

Solution

a)

$$\begin{aligned}
E[V] &= E[E(V|Y)] \text{by the law of iterated expectations} \\
&= E[E(Q - \tilde{Q}|Y)] \\
&= E[E(Q|Y) - E(\tilde{Q}|Y)] \\
&= E[E(Q|Y) - E(E[Q|Y]|Y)] \\
&= E[E(Q|Y) - E(Q|Y)] = 0
\end{aligned}$$

b)

$$\begin{aligned}
E[VY] &= E[E(VY|Y)] \text{ by LIE} \\
&= E[YE(V|Y)] \text{by properties of conditional expectations } E(YV|Y) = YE(V|Y) \\
&= E[Y \times 0] = 0 \text{ by the a)}
\end{aligned}$$

c)

Following the hints let

$$h(Y) = g(Y) - E[Q|Y]$$
$$W = V - h(Y) = (Q - E[Q|Y]) - h(Y)$$

Let us begin from

$$\begin{aligned} E[W^2] &= E[(V - h(Y))^2] \\ &= E[V^2] - 2E[Vh(Y)] + E[h(Y)^2] \text{ by linearity of expectations} \\ &= E[V^2] - 2 \times 0 + E[h(Y)^2] \text{ similarly as in b)} \\ &= E[V^2] + E[h(Y)^2] \end{aligned}$$

Now we have

$$E[W^2] \geq E[V^2] \implies E[V^2] + E[h(Y)^2] \geq E[V^2]$$

which holds true if the $E[h(Y)^2]$ is positive.

We have that

$$E[h(Y)^2] = Var(h(Y)) + (E[h(Y)])^2$$

which is positive. Thus we have shown that

$$E[W^2] \geq E[V^2].$$

Exercise 8

(18p) In this exercise you prove the unbiasedness of the population variance. You have an iid sample $Y_1, ..., Y_n$ with mean μ_Y and variance σ_Y^2. Remember to state what assumptions or properties you use and where.

-Show that $E[(Y_i - \overline{Y})^2] = Var(Y_i) - 2cov(Y_i, \overline{Y}) + Var(\overline{Y})$.

-Show that $cov(\overline{Y}, Y_i) = \frac{\sigma_Y^2}{n}$

-Show that $E(s_Y^2) = \sigma_Y^2$ using the results you obtained in (a) and (b)

Hints: (a) Linearity of Expectation: $E[Y+X] = E[Y]+E[X]$. (b) A property of variance: $Var(a+bY) = b^2\sigma_Y^2$. What does the fact that the observations are iid imply for the covariance?

Solution

a)

$$\begin{aligned} E[(Y_i - \overline{Y})^2] &= E[((Y_i - \mu_Y) - (\overline{Y} - \mu_Y))^2] \text{ add and subtract } \mu_Y \\ &= E[(Y_i - \mu_Y)^2] - 2(Y_i - \mu_Y)(\overline{Y} - \mu_Y) + (\overline{Y} - \mu_Y)^2] \\ &= E[(Y_i - \mu_Y)^2] - 2E[(Y_i - \mu_Y)(\overline{Y} - \mu_Y)] + E[(\overline{Y} - \mu_Y)^2] \text{ LIE*} \\ &= Var(Y_i) - 2cov(Y_i, \overline{Y}) + Var(\overline{Y}) \text{ definitions of variance and covariance} \end{aligned}$$

*Linearity of expectations

b)

Keeping in mind that

$$\frac{1}{n}\sum_{j=1}^{n} Y_i - \mu_Y = \frac{1}{n}(Y_1 + ... + Y_n) - \mu_Y = \frac{1}{n}(Y_1 + ... + Y_n - n\mu_Y)$$

$$\begin{aligned}
cov(\overline{Y}, Y_i) &= E[(\overline{Y} - \mu_Y)(Y_i - \mu_Y)] \\
&= E\left[\left(\frac{1}{n}\sum_{j=1}^{n} Y_j - \mu_Y\right)\left(Y_i - \mu_Y\right)\right] \\
&= E\left[\left(\frac{\sum_{j=1}^{n} Y_j - \mu_Y}{n}\right)\left(Y_i - \mu_Y\right)\right] \text{ because } \sum_{j=1}^{n} \mu_Y = n\mu_Y \\
&= \frac{1}{n}E\left[\left((Y_1 - \mu_Y) + ... + (Y_i - \mu_Y) + ... + (Y_n - \mu_Y)\right)\left(Y_i - \mu_Y\right)\right] \\
&= \frac{1}{n}(cov(Y_1, Y_i) + ... + cov(Y_i, Y_i) + ... + cov(Y_n, Y_i)) \\
&= \frac{1}{n}cov(Y_i, Y_i) \text{ because observations are iid, } cov(Y_j, Y_i) = 0 \text{ for } j \neq i \\
&= \frac{1}{n}Var(Y_i) = \frac{\sigma_Y^2}{n}
\end{aligned}$$

c)

$$\begin{aligned}
E(s_Y^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2\right] = \frac{1}{n-1}\sum_{i=1}^{n} E[(Y_i - \overline{Y})^2] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[Var(Y_i) - 2cov(Y_i, \overline{Y}) + Var(\overline{Y})] \text{ by a)} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sigma_Y^2 - 2\frac{\sigma_Y^2}{n} + Var(\overline{Y})] \text{ by b)} \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sigma_Y^2 - 2\frac{\sigma_Y^2}{n} + \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i)] \text{ by } Var(aY) = a^2Var(Y) \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sigma_Y^2 - 2\frac{\sigma_Y^2}{n} + \frac{1}{n^2}\sum_{i=1}^{n}\sigma_Y^2] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\sigma_Y^2 - 2\frac{\sigma_Y^2}{n} + \frac{1}{n^2}n\sigma_Y^2] = \frac{1}{n-1}\sum_{i=1}^{n}[\sigma_Y^2 - \frac{\sigma_Y^2}{n}] \\
&= \frac{1}{n-1}\sum_{i=1}^{n}[\frac{n\sigma_Y^2 - \sigma_Y^2}{n}] = \frac{1}{n-1}(n-1)\sigma_Y^2 = \sigma_Y^2
\end{aligned}$$