W203-1 - Fall 2014 - Lab 2

Juan Jose Carin

Sunday, October 19, 2014

Contents

Part 1: Multiple Choice **3**

1) a) Bar charts . . . 3

2) c) $H_0 : \mu = \mu_0; H_a : \mu > \mu_0$ . . . 3

3) f) None of the above . . . 3

4) e) a and d . . . 3

5) e) Raise the variable to a power greater than 1 . . . 3

6) b) The standard deviation of Berkeley student ages is 2 years . . . 3

7) d) What is the probability of the data we observe, assuming that the null hypothesis is true? . . 3

8) c) Assuming your null hypothesis is actually false, your p-value is likely to decrease as you increase your sample size . . . 3

9) d) Independence of observations . . . 4

10) f) None of the above . . . 4

Part 2: Test Selection **4**

1) b) Levene's test . . . 4

2) a) Shapiro-Wilk test . . . 4

Part 3: Data Analysis and Short Answer **4**

1) Data Import and Error Checking . . . 4

a. Examine the agewed variable (age when married). . . . 4

i. What are the value(s) of agewed, if any, that do not meaningfully correspond to ages? 6

b. Recode any value(s) that do not correspond to age as NA. . . . 7

i. What is the mean of the agewed variable? . . . 7

2) Checking assumptions . . . 8

a. Produce a QQ plot for the agewed variable. . . . 8

i. Using this plot information, is agewed normal and how do you know? . . . 9

b. Perform a Shapiro-Wilk test on the agewed variable. . . . 9

i. What is the null and alternative hypothesis for your test? . . . 9

ii. What is your p-value, and what is your specific conclusion? . . . 9

c. What is the variance of agewed for men? What is the variance of agewed for women? . . . 10

d. Perform a Levene's test for the agewed variable grouped by men and women. . . . 10

i. What is the null and alternative hypothesis for this test? 11

ii. What is your p-value, and what is your specific conclusion? 11

3. More hypothesis testing 11

a. Suppose we have a hypothesis that the age of marriage (agewed) in the population has a mean of exactly 23, with a standard deviation of 5 years (you should assume this value is correct rather than estimating the standard deviation from the data). Perform a z-test to analyze this hypothesis. 11

i. What is the null and alternative hypothesis for this test? 12

ii. What is your p-value, and what is your specific conclusion? 12

Part 1: Multiple Choice

I've included comments in some sections of this Part, (not—only—to justify my choice but) to document my own reasoning when selecting the right answer, for future reference.

1) a) Bar charts

Comment: *Line charts would suggest a sequence (a time series), which discards (e) and (f).*

2) c) $H_0 : \mu = \mu_0; H_a : \mu > \mu_0$

Comment: *As usual, the null hypothesis is an equality. The alternative hypothesis involves a specific direction (>), which in turn involves a one-tailed test, because we're interested in testing the claim that* μ *is greater than* $\mu_0 = 10$.

3) f) None of the above

Comment: *The p value alone tells us nothing about neither alpha nor the effect size.*

4) e) a and d

(The possiblity of a Type II error will go up and the statistical power will go down in the second study.)

5) e) Raise the variable to a power greater than 1

Comment: *To correct the negative skew.*

6) b) The standard deviation of Berkeley student ages is 2 years

Comment: *Since the null hypothesis is usually an equality statement, we discard (a), which involves a range, and (c) and (d), which involve inequalities.*

*The selected null hypothesis could be tested, for instance, using a Chi-square test for variance (**provided that** the observations are independent and come from a normal distribution).*

7) d) What is the probability of the data we observe, assuming that the null hypothesis is true?

8) c) Assuming your null hypothesis is actually false, your p-value is likely to decrease as you increase your sample size

Comment: *(f) (Assuming your null hypothesis is actually true, and you were to repeat the experiment a large number of times, you would expect a type 1 error 4% of the time) seemed a little bit tricky, but 4% is the value of type I error for the specific sample that has been drawn in this particular study. Subsequent samples would lead to a different sample mean and hence to different test statistics and p-values (=type I errors).*

9) d) Independence of observations

Comment: *(d) Since participants are rewarded if they correctly recall, and they can complete the experiment during a really long period, it is likely that some of them get in contact and influence each other's responses.*

10) f) None of the above

Comment: *(d) makes no sense at all, and the rest involve Type I or Type II errors, or statistical power, and the three of them imply certainty about which is true (type I error implies that it's the null hypothesis which is true, and type II error and power imply that the alternative is the right one.*

Part 2: Test Selection

1) b) Levene's test

2) a) Shapiro-Wilk test

Part 3: Data Analysis and Short Answer

1) Data Import and Error Checking

The GSS dataframe contains information about 1500 people regarding to 46 variables (plus their ID number): age, month of birth, work & marital status, number of siblings and children, race, income…

a. Examine the agewed variable (age when married).

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   18.00   21.00   19.06   24.00   99.00
```

```
##  nbr.val nbr.null   nbr.na      min      max    range      sum
##     1500      286        0        0       99       99    28584
```

None of the 1500 observations in the GSS dataframe have NA values for the agewed variable. The variable ranges from 0 to 99, the Interquantile Range is 24-18 years (i.e., according to the GSS dataframe 50% of the 1500 people contained in it got married within that range of ages), the median is 21 years, and the mean is 19.1 years.

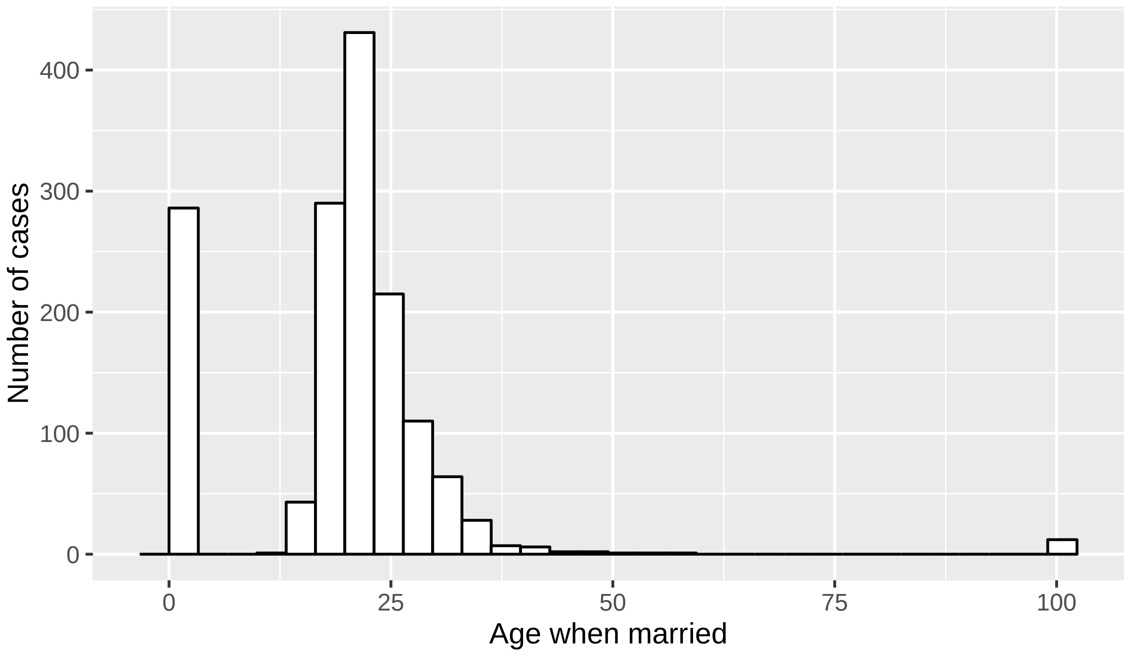

Figure 1: Histogram of the agewed variable

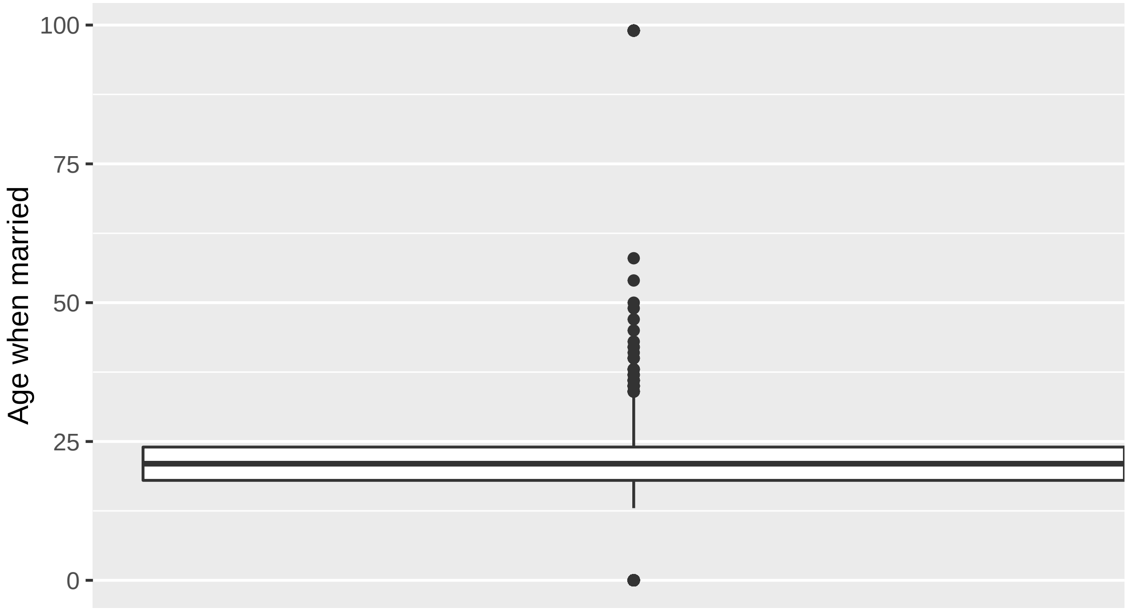

Figure 2: Boxplot of the agewed variable

i. What are the value(s) of agewed, if any, that do not meaningfully correspond to ages?

1) The minimum value of the agewed variable is **0**… and—as shown below—that's the value for 286 of the cases (19.1% of the total). It seems very unlikely that newborns (whether so many or even a few) get married. Moreover, if we analyze the marital status of those people, we find out that all of them have never been married, so the problem is that the variable was unproperly coded in their case.

```
  # Check number of cases of agewed=0 and associated marital status
sum(GSS$agewed == 0)
```

```
## [1] 286
```

```
table(GSS$marital, GSS$agewed==0)[, 2]
```

```
##       married       widowed      divorced     separated never married 
##             0             0             0             0           286 
##            NA 
##             0
```

2) The maximum value of the agewed variable (**99**) is unlikely but possible. But looking at the histogram on the previous page (*Figure 1*) we see that there are a significant number of cases with a similar value, so we check high values (based on the histogram, let's say greater than 90) of the variable, and find out—as shown below—that there are 12 people whose value of the agewed variable is exactly 99 (0.8 % of the total), which does not seem very probable. This may be due to an unproper coding, or to errors entering the data, which should be corrected anyway.

```
  # Check number of cases of agewed>90, the exact value for all those cases
  # and associated marital status
sum(GSS$agewed > 90)
```

```
## [1] 12
```

```
GSS[GSS$agewed > 90, "agewed"]
```

```
##  [1] 99 99 99 99 99 99 99 99 99 99 99 99
```

```
table(GSS$marital, GSS$agewed == 99)[, 2]
```

```
##       married       widowed      divorced     separated never married 
##             2             2             6             1             0 
##            NA 
##             1
```

b. Recode any value(s) that do not correspond to age as NA.

We just have to assign NA to those cases where agewed is **0** or **99**.

```
GSS$agewed[GSS$agewed == 0 | GSS$agewed == 99] <- NA
```

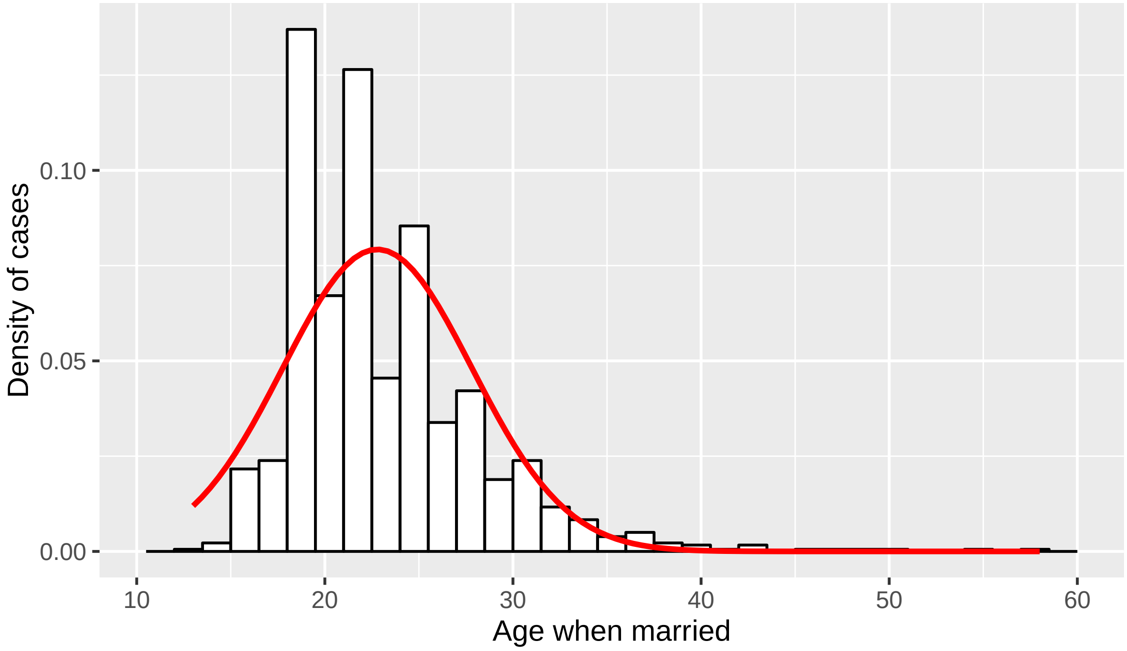

Figure 3: New histogram of the agewed variable

i. What is the mean of the agewed variable?

Now (having 298 NA cases) the agewed variable ranges from 13 to 58, the Interquantile Range becomes 25-19 years, the median is 22 years, and **the mean is 22.8 years**.

```
   # New (and valid) mean
agewed.mean <- mean(GSS$agewed, na.rm = TRUE)
print(paste("The mean of the agewed variable is",
            format(agewed.mean, digits=4), "years"), sep=" ")
```

```
## [1] "The mean of the agewed variable is 22.79 years"
```

2) Checking assumptions

a. Produce a QQ plot for the agewed variable.

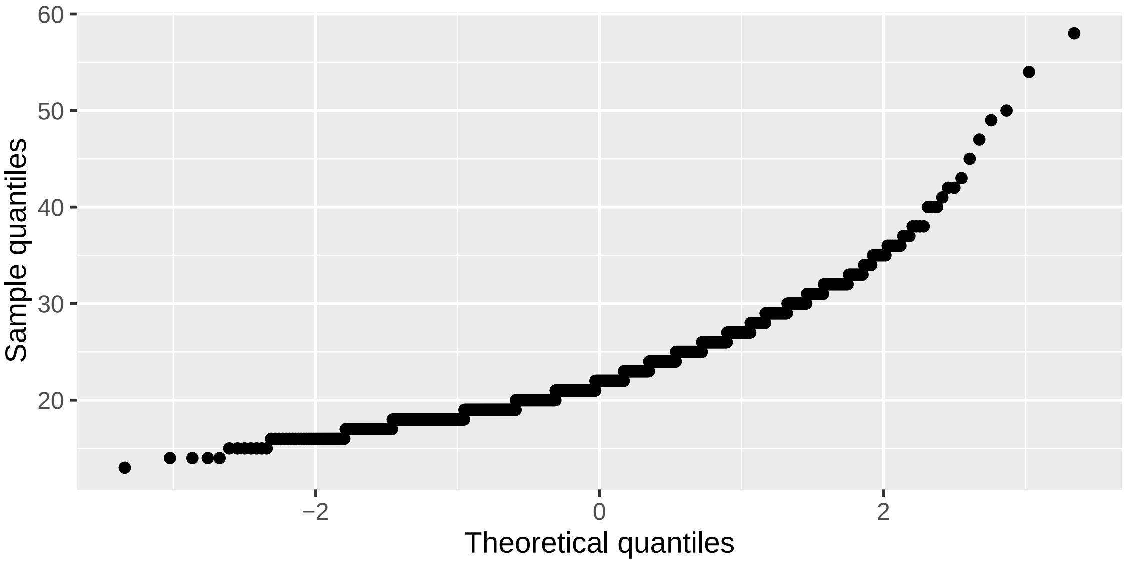

Figure 4: QQ-plot of the agewed variable

Though the plot above is quite explanatory, we may want to standardize the agewed variable. This way we are able to better compare the cumulative values of this sample to those of the standard normal distribution *N(0,1)* (as we know, if the sample distribution were normal, the plot would be very similar to the straight diagonal line y=x).

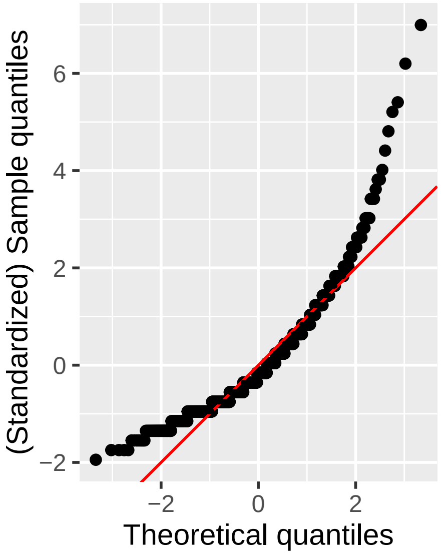

Figure 5: QQ-plot of the agewed variable standardized

i. Using this plot information, is agewed normal and how do you know?

(As mentioned in the previous page) If agewed were normal, the QQ plot would have shown a straight diagonal line. This is not the case, so **agewed is far from being normal**. To give a more detailed explanation:

1) The left tail twists off clockwise from the reference line (so there is less data in the left tail than in a normal distribution), and the right tail twists counterclockwise (so there is more data in the right tail than in a normal distribution). Both facts, combined, indicate **positive skew**.
2) The fact that the center region of the plot lies below the diagonal line indicates **positive kurtosis** or **leptokurtosis**.

b. Perform a Shapiro-Wilk test on the agewed variable.

```
shapiro.test(GSS$agewed)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  GSS$agewed
## W = 0.88959, p-value < 2.2e-16
```

```
shapiro.test(GSS$agewed)$p.value
```

```
## [1] 1.816354e-28
```

i. What is the null and alternative hypothesis for your test?

The **null-hypothesis** is that the values or scores of the agewed variable are normally distributed, and the **alternative hypothesis** is that they are not

ii. What is your p-value, and what is your specific conclusion?

As seen above, **the p-value is 1.8e-28**. Since it is much smaller than any α level we could have chosen, **the null hypothesis is rejected**—we have highly statistically significant evidence that the agewed scores are not from a normally distributed population.

Anyway, we should consider that our sample is very large (1202 scores), so the result of the Shapiro-Wilk test may not be considered conclusive. **The conclusion above comes also from the previous QQ plots and histogram** (*Figures 3* to 5).

We could also consider the specific values of kurtosis and skew:

```r
stat.desc(GSS$agewed, basic = FALSE, desc = FALSE, norm = TRUE)
```

```
##     skewness     skew.2SE     kurtosis     kurt.2SE   normtest.W
## 1.653714e+00 1.171786e+01 5.340543e+00 1.893660e+01 8.895862e-01
##   normtest.p
## 1.816354e-28
```

Since the values of skew and kurtosis divided by 2 standard errors are, respectively, 11.7 and 18.9, both much greater than $3.29/2 \simeq 1.65$, both are highly significant (at p<0.001)... but again, the sample size has a great impact on those values.

c. What is the variance of agewed for men? What is the variance of agewed for women?

```r
by(GSS$agewed, GSS$sex, var, na.rm = TRUE)
```

```
## GSS$sex: Male
## [1] 23.6843
## ------------------------------------------------------------
## GSS$sex: Female
## [1] 24.29948
```

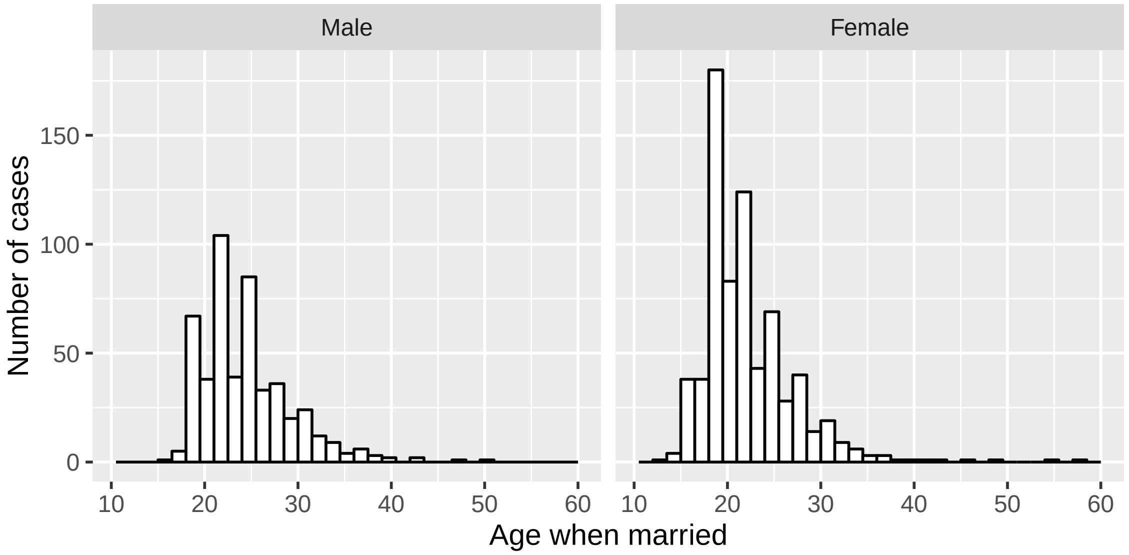

Figure 6: Histogram of the agewed variable by sex

d. Perform a Levene's test for the agewed variable grouped by men and women.

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  0.9609 0.3272
##       1200
```

i. What is the null and alternative hypothesis for this test?

The **null-hypothesis** is that the variances of a variable through different groups or levels of another variable are equal, i.e., there is homogeneity of variance. The **alternative hypothesis** is that there is a difference between the variances of each group.

ii. What is your p-value, and what is your specific conclusion?

As seen above, **the p-value is 0.327**. Since it is much greater than 0.05, it is not statistically significant, and hence **we fail to reject the null hypothesis—we retain it**.

Anyway, we should consider again that our sample is very large (1202 scores), so the result of the Levene's test may not be considered conclusive.

As a double check, we could look at **Hartley's** F_max or **variance ratio**: for such a large sample, the value of the test statistic F must be smaller than 1 (regardless of the number of variances being compared) to be non-significant. Because the value we have obtained is *F(1,1200) = 0.961 < 1* **we can maintain our previous conclusion that the result is non-significant**.

3. More hypothesis testing

a. Suppose we have a hypothesis that the age of marriage (agewed) in the population has a mean of exactly 23, with a standard deviation of 5 years (you should assume this value is correct rather than estimating the standard deviation from the data). Perform a z-test to analyze this hypothesis.

```r
mu0 <- 23
SD <- 5
N <- sum(!is.na(GSS$agewed))
SE <- SD/sqrt(N)
sample.mean <- mean(GSS$agewed, na.rm = TRUE)
z <- (sample.mean - mu0) / SE
z
```

```
## [1] -1.442174
```

```r
p.value <- (1 - pnorm(abs(z)))*2
p.value
```

```
## [1] 0.1492532
```

We could also make use of the *z.test* function in the *BSDA* package:

```r
library(BSDA)
z.test(GSS$agewed[!is.na(GSS$agewed)], alternative = "two.sided", mu = mu0,
       sigma.x = SD, conf.level = 0.95)
```

```
##
##  One-sample z-Test
##
## data:  GSS$agewed[!is.na(GSS$agewed)]
## z = -1.4422, p-value = 0.1493
## alternative hypothesis: true mean is not equal to 23
## 95 percent confidence interval:
##  22.50935 23.07467
## sample estimates:
## mean of x
##  22.79201
```

i. What is the null and alternative hypothesis for this test?

The **null-hypothesis** is that the mean value of the agewed variable for the whole population (not for the sample included in the GSS dataframe) is the mentioned $\mu_0 = 23$. The **alternative hypothesis** is that the mean of the population is different (greater or smaller; i.e., this is a two-tailed test).

The assumption in a z-test, under the null hypothesis, is that the distribution of the test statistic can be approximated by a normal distribution. In this particular case the test statistic is the sample mean ($barX$), and because of the central limit theorem and our sample size (1202 scores, excluding NA values), we can state that this assumption is true (i.e., $\bar{X}$ will be approximately normally distributed). And so will be the corresponding z score ($z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{N}}$), with a mean of 0 and a standard deviation of 1.

To calculate the standard error—and hence the z score—we need to know the standard deviation of the population (σ). In this particular case we know that $\sigma = 5$; otherwise we would have to use an approximate value, e.g., the standard deviation of the sample (because our sample is so large, we still could use the z-test in that case; otherwise a t-test would be more appropriate to account for the uncertainty in the sample variance).

ii. What is your p-value, and what is your specific conclusion?

The p-value for this two-tailed test is calculated as $2\Phi(-|z|)$ or $2(1-\Phi(|z|)$, where Φ is the standard normal cumulative distribution function (function *pnorm* of **R**).

As seen above, z is -1.44 and **the p-value is 0.149**. That means that, under the null hypothesis, the probability of a simple random sample of 1202 people having a mean not as extreme as the one contained in the GSS dataframe—i.e., larger than 22.79 or smaller than 23.21—would be "only" 0.851. In other words, since the p-value is greater than 0.05, it is not statistically significant, and hence **we fail to reject the null hypothesis, and we retain it**.