Chapter 7 - Inference for Numerical Data

**Working backwards, Part II.** (5.24, p. 203) A 90% confidence interval for a population mean is (65, 77). The population distribution is approximately normal and the population standard deviation is unknown. This confidence interval is based on a simple random sample of 25 observations. Calculate the sample mean, the margin of error, and the sample standard deviation.

**ANSWER**

**Sample Mean:**

As we know, the sample mean is the average of the sample values. I calculated it by summing all observations and dividing by the number of observations.

Values that are given:-  
X2 = 77 and X1 = 65

Sample Mean = (X2+X1) = (77 + 65)/2 = 71

**Margin of Error (ME):**

The margin of error indicates the range within which the true population mean lies, with a certain level of confidence. I calculated it as the difference between the two extreme sample values divided by 2:

Margin of Error = (X2-X1)/2 = (77−65)/2 ​= 6

**Degrees of Freedom (df):**

Degrees of freedom are typically associated with a sample size n. For a t-distribution, the degrees of freedom are calculated as n−1. In this case, with n=25, the degrees of freedom are:

df = n−1 = 25−1 = 24

Using the degrees of freedom, I look up the t-value for a 90% confidence interval (or 95%, depending on the assumption). In this case, for a 90% confidence interval:

Tcritical = t\_0.95(df=24)

This value is approximately t\_(0.95) = 1.711.

**Standard Error (SE):**

We can calculate the standard error by dividing the margin of error by the critical value of the t-distribution, below is the formula:

Standard Error (SE) = Margin of Error/t\_critical

Substituting the values:

SE = 6/1.711 ≈ 3.507

**Sample Standard Deviation (sd):**

We can calculate the sample standard deviation by using the relationship between the standard error and the sample size n, below is the formula:

Standard Deviation(σ) = SE × sqrt(n)​

Substituting the values:

σ = 3.507 × sqrt(25) = 3.507×5 = 17.535

Final Results:

* Sample Mean xˉ = 71
* Margin of Error ME = 6
* Sample Standard Deviation σ = 17.535

**SAT scores.** (7.14, p. 261) SAT scores of students at an Ivy League college are distributed with a standard deviation of 250 points. Two statistics students, Raina and Luke, want to estimate the average SAT score of students at this college as part of a class project. They want their margin of error to be no more than 25 points.

1. Raina wants to use a 90% confidence interval. How large a sample should she collect?
2. Luke wants to use a 99% confidence interval. Without calculating the actual sample size, determine whether his sample should be larger or smaller than Raina’s, and explain your reasoning.
3. Calculate the minimum required sample size for Luke.

**ANSWER**

**Formula for Sample Size (n):**

We can calculate the sample size required to estimate a population mean with a given confidence level by the following below formula:

N = (z∗ x σ /E)^2

Where:

* n is the required sample size.
* z∗ is the z-score corresponding to the desired confidence level.
* σ is the standard deviation of the population.
* E is the margin of error (desired precision).

In your case:

* σ=250 (given population standard deviation).
* E=25 (desired margin of error).
* z∗ varies based on the confidence level (90% or 99%).

**A. Raina's Sample Size with 90% Confidence Interval:**

For Raina, the confidence level is 90%, and the corresponding z-score will be z∗ = 1.645

Using below formula for sample size:

N = (1.645 x 250/25)^2

First, I will calculate the expression inside the parenthesies:

(1.645 x 250)/25 = 411.25/25 = 16.45

Now squaring above result:

N = (16.45)^2 = 270.60

Since sample size must be a whole number, I’ve rounded this to the nearest whole number that is:

n=271

Thus, Raina might collect a sample size of **271**.

**B. Luke's Sample Size with 99% Confidence Interval:**

Without doing the calculations, We know Luke's confidence level is higher (99% confidence), so the corresponding z-score is larger. This implies that Luke’s sample size should be **larger** than Raina’s because a higher confidence level requires a wider range of z-values, leading to a larger sample to achieve the same margin of error.

**C. Luke's Sample Size with 99% Confidence Interval:**

For Luke, the confidence level is 99%, and the corresponding z-score will be z∗ = 2.576

Using the same sample size formula as we used above:

n=(2.576 x 250/25)^2

First, I will calculate the expression inside the parenthesies:

2.576 x 250/25 = 644/25

Now squaring above result:

N = (25.76)^2=663.78

Rounding this to the nearest whole number:

n=664

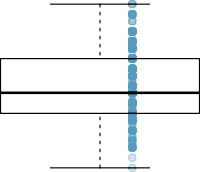
Thus, Luke might collect a sample size of **664**.

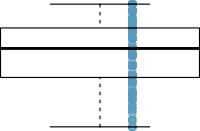
**Final Result’s:**

* **Raina’s sample size** (90% confidence): **271**
* **Luke’s sample size** (99% confidence): **664**

Since Luke’s confidence level is higher, his required sample size is larger than Raina’s, as expected.

**High School and Beyond, Part I.** (7.20, p. 266) The National Center of Education Statistics conducted a survey of high school seniors, collecting test data on reading, writing, and several other subjects. Here we examine a simple random sample of 200 students from this survey. Side-by-side box plots of reading and writing scores as well as a histogram of the differences in scores are shown below.

80 40

30

60

scores

20

40 10

20 0

read write

−20 −10 0 10 20

Differences in scores (read − write)

1. Is there a clear difference in the average reading and writing scores?
2. Are the reading and writing scores of each student independent of each other?
3. Create hypotheses appropriate for the following research question: is there an evident difference in the average scores of students in the reading and writing exam?
4. Check the conditions required to complete this test.
5. The average observed difference in scores is *x*^*read−write* = *−*0*.*545, and the standard deviation of the differences is 8.887 points. Do these data provide convincing evidence of a difference between the average scores on the two exams?
6. What type of error might we have made? Explain what the error means in the context of the application.
7. Based on the results of this hypothesis test, would you expect a confidence interval for the average difference between the reading and writing scores to include 0? Explain your reasoning.

**ANSWER**

1. No. The averages appear to differ slightly, but the pattern of the differences seems to follow a normal distribution only.
2. No. We can see the reading and writing scores are independent of each other.
3. **Null Hypothesis (H₀): HO: ur – uw equal to 0.** The mean reading and writing scores are equal, indicating no difference between them.

**Alternative Hypothesis (Hₐ): HA: ur – uw not equal to 0.** The mean reading and writing scores are not equal, suggesting a difference between them.

1. **Independence of observations:** The difference histogram indicates that the data are paired, meaning the observations are not independent.

**Nearly normal distribution of observations:** The box plot in the text shows that the data appear to be reasonably normally distributed, with no significant outliers.

1. **Standard Error (SE):**SE = 8.887/sqrt(200) = 0.628

**t-statistic:**

t = (−0.545−0)/SE = −0.545/0.628 = −0.868

**p-value:**

p = 2 × P (t\_199 ≤ −0.868) = 0.386

**Conclusion:**

The computed p-value is 0.386, which leads us to fail to reject the alternative hypothesis (HA​). This suggests there is no strong evidence to support a difference in the average means.

1. By failing to reject HA, if HA was correct, we would have committed a Type II error. This implies that while I accepted H0, there truly is a difference in the means. Increasing the sample size would reduce the likelihood of making a Type II error.
2. We would anticipate the confidence intervals to contain 0, since there is no strong evidence suggesting a difference in the average means.

**Fuel efficiency of manual and automatic cars, Part II.** (7.28, p. 276) The table provides summary statistics on highway fuel economy of cars manufactured in 2012. Use these statistics to calculate a 98% confidence interval for the difference between average highway mileage of manual and automatic cars, and interpret this interval in the context of the data.

15

25

Hwy MPG

Automatic Manual Mean 22.92 27.88

SD 5.29 5.01

n 26 26

35

# automatic manual

Hwy MPG

**ANSWER**

**Hypothesis Definition:**

Null Hypothesis (H₀):  
There is no difference between the average miles per gallon (MPG) for automatic and manual transmissions. H0: μa−μm is equal to 0

Alternative Hypothesis (Hₐ):  
There is a difference between the average MPG for automatic and manual transmissions.

HA:μa−μm is not equal to 0

**Set the Significance Level (α)**

The significance level is given as:

Α = 0.05

**Calculating the Mean Difference now (mdiff)**

The sample mean difference between automatic (μa​) and manual (μm) transmissions is calculated below:

Mdiff = μa−μm = 16.12−19.85 = −3.73

**Calculating the Standard Error of the Difference (SEdiff)**

To calculate the standard error of the difference between the two sample means, we can use the below formula:

SEdiff = sqrt [ (σa^2/n) + (σm^2/n) ]

Where:

* σa=3.58 (standard deviation for automatic)
* σm=4.51(standard deviation for manual)
* n=26 (sample size)

Substituting the values:

SEdiff = sqrt [ (3.58)^2/26 + (4.51)^2/26 ]= sqrt [ 12.8164/26 + 20.3401/26 ]= sqrt [0.4929 + 0.7823] = sqrt[1.2752] = 1.129

**Calculating the t-Statistic (t) now**

The t-statistic is calculated using the below formula:

t = mdiff−0/SEdiff = (−3.73−0)/1.129 = −3.303

**Calculating the p-Value (p)**

To calculate the p-value, we will use the t-distribution with n−1 degrees of freedom, where n=26:

P = 2 × P(t\_25 ≤ −3.303) = 0.003

Using statistical tables or software, I found:

p approximately 0.003

**Conclusion**

With α = 0.05 and p = 0.003, the p-value is less than the significance level:

p < α

Thus, we are good to reject the null hypothesis H0

**So we can say,** there is convincing evidence to suggest a difference in the average miles per gallon between manual and automatic transmissions.

**Email outreach efforts.** (7.34, p. 284) A medical research group is recruiting people to complete short surveys about their medical history. For example, one survey asks for information on a person’s family history in regards to cancer. Another survey asks about what topics were discussed during the person’s last visit to a hospital. So far, as people sign up, they complete an average of just 4 surveys, and the standard deviation of the number of surveys is about 2.2. The research group wants to try a new interface that they think will encourage new enrollees to complete more surveys, where they will randomize each enrollee to either get the new interface or the current interface. How many new enrollees do they need for each interface to detect an effect size of 0.5 surveys per enrollee, if the desired power level is 80%?

**ANSWER**

**Below is the formula for Sample Size:**

N = (Z x σE)^2

Where:

* Z=1.28 (Z-score for 80% confidence level)
* σ=2.2 (Standard deviation)
* E=0.5 (Margin of error)

Substituting the values:

N = (1.28 x 2.2/0.5)^2

First, we will calculate the expression inside the parentheses:

N = (2.816/0.5)^2 = (5.632)^2

Now, squaring the result:

n=31.72

Thus, the required sample size is approximately 32.

**Work hours and education.** The General Social Survey collects data on demographics, education, and work, among many other characteristics of US residents.47 Using ANOVA, we can consider educational attainment levels for all 1,172 respondents at once. Below are the distributions of hours worked by educational attainment and relevant summary statistics that will be helpful in carrying out this analysis.

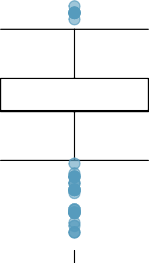
*Educational attainment*

Less than HS HS Jr Coll Bachelor’s Graduate Total Mean 38.67 39.6 41.39 42.55 40.85 40.45

SD 15.81 14.97 18.1 13.62 15.51 15.17

n 121 546 97 253 155 1,172

80



Hours worked per week

60

40

20

0

Less than HS HS Jr Coll Bachelor's Graduate

1. Write hypotheses for evaluating whether the average number of hours worked varies across the five groups.
2. Check conditions and describe any assumptions you must make to proceed with the test.
3. Below is part of the output associated with this test. Fill in the empty cells.

Df Sum Sq Mean Sq F-value Pr(*>*F)

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| degree  Residuals |  |  |  | 501.54 | | |  | 0.0682 |
|  | 267,382 | | |  |  | | |
| Total |  |  |  |  | | | | |

1. What is the conclusion of the test?

**ANSWERS**

**A.**

**Null Hypothesis (H₀):**  
The average hours worked are equal across all five groups, indicating that there is no variation.

μHS=μJR=μB=μG

**Alternative Hypothesis (Hₐ):**  
The average hours worked differ among some or all of the groups, suggesting that there is variation.  
  
**B.**   
The data for each group seems to follow a roughly normal distribution, and the variation within each group is comparable. To conduct an ANOVA test, we can assume that the observations are independent.

**C**.

**1. Degrees of Freedom (df):**

* **Groups (df\_G)**:

ff\_G = k−1 = 5−1 = 4

* **Error (df\_E)**:

df\_E = n−k = 1172−5 = 1167

* **Total (df\_T)**:

df\_T = df\_G + df\_E = 4 + 1167 = 1171

**2. Sum of Squares Within (SSW):**

To calculate the Sum of Squares Within (SSW), we sum the squared deviations within each group. This is done by multiplying the variance within each group by n−1 (the degrees of freedom for each group):

SSW = 5∑i=1 ((ni−1) x (si)^2)

SSW = (121−1) x (15.81)^2 + (546−1) x (14.97)^2 + (97−1) x (18.1)^2 + (253−1) x (13.62)^2 + (155−1) x (15.51)^2

SSW = 120 x 249.6961 + 545 x 224.1009 + 96 x 327.61 + 252 x 185.3444 + 154 x 240.9601

SSW = 29963.53 + 122134.99 + 31450.56 + 46606.79 + 37008.86

SSW = 267165 (rounded to nearest whole number)

**3. Sum of Squares Between (SSB):**

To calculate the Sum of Squares Between (SSB), we sum the squared deviations of each group's mean from the grand mean, multiplied by the group size:

SSB = 5∑i=1 (ni x (xˉI − xˉ)^2)

Substituting the values (group sizes and means):

SSB=121 x (38.67−40.45)^2 + 546 x (39.60−40.45)^2 + 97 x (41.39−40.45)^2 + 253 x (42.55−40.45)^2 + 155 x (40.85−40.45)^2.

Calculating:

SSB = 121 x (−1.78)^2 + 546 x (−0.85)^2 + 97 x (0.94)^2 + 253 x (2.10)^2 + 155 x (0.40)^2

SSB = 121 x 3.1684 + 546 x 0.7225 + 97 x 0.8836 + 253 x 4.41 + 155 x 0.16

SSB=121 x 3.1684 + 546 x 0.7225 + 97 x 0.8836 + 253 x 4.41 + 155 x 0.16

SSB = 383.37 + 394.89 + 85.51 + 1115.73 + 24.8

SSB=2004 (rounded to the nearest whole number)

**4. Total Sum of Squares (SST):**

The Total Sum of Squares (SST) is the sum of the within-group and between-group sums of squares:

SST = SSW+SSB = 267165 + 2004 = 269169

**5. Mean Squares Between (MSB):**

The Mean Square Between (MSB) is calculated by dividing the sum of squares between by its degrees of freedom:

MSB = SSB/df\_G = 2004/4 = 501

**6. Mean Squares Within (MSW):**

The Mean Square Within (MSW) is calculated by dividing the sum of squares within by its degrees of freedom:

MSW = SSW/df\_E = 267165/1167 = 228.86

**7. F-statistic (F):**

The F-statistic is calculated by dividing the mean square between by the mean square within:

F = MSB/MSW = 501/228.86 ≈ 2.189

**Final Results:**

* **Degrees of Freedom Between (df\_G):** 4
* **Degrees of Freedom Within (df\_E):** 1167
* **Sum of Squares Between (SSB):** 2004
* **Sum of Squares Within (SSW):** 267165
* **Total Sum of Squares (SST):** 269169
* **Mean Square Between (MSB):** 501
* **Mean Square Within (MSW):** 228.86
* **F-statistic:** 2.189

**D.** There isn't sufficient evidence to reject the null hypothesis, so we conclude that there is no strong indication that the average hours worked per week differ across the five educational groups.