Rohit Wason
Spring 2021

Math 560 Homework (#7, Inference of the Mean)

Problem 1. Given $\bar{x} = 9.289221, s = 0.8191156$ and $n = 10$ calculate the 92% two-sided C.I. for μ.

Solution.

- t^*, a statistic that follows a **t-distribution**, $t(n-1=7)$ for $C = 92\%$, is given by

```
qt(0.96,df=7)
>> 2.046011
```

- Hence the 92% confidence interval for μ is:

$$
\begin{aligned}
&= \bar{x} \pm t^* \frac{s}{\sqrt{n}} \\
&= 9.289221 \pm 2.046011(\frac{0.8191156}{\sqrt{8}}) \\
&= (8.696694, 9.881748)
\end{aligned}
$$

□

Problem 2. StatsVillage.txt is used.

Solution. (a) The following seed is used

```
set.seed(19891)
```

(b) Here are the summary statistics for two independent SRSs

Population	Name	n	$\bar{x}$	s
1	North	15	3474.8	7828.715
2	South	20	7319.15	5318.922

1. Test $H_0 : (\mu_1 - \mu_2) = 0$ vs. $H_a : (\mu_1 - \mu_2) < 0$

2. The test statistic:

$$\begin{aligned} t &= \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \\ &= \frac{(3474.8 - 7319.15) - 0}{\sqrt{\frac{7828.715^2}{15} + \frac{5318.922^2}{20}}} \\ &= -1.639167 \end{aligned}$$

3. For critical value, t^* first let's calculate degrees of freedom:

$$\begin{aligned} k &= \frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{1}{n_1-1}(\frac{s_1^2}{n_1})^2 + \frac{1}{n_2-1}(\frac{s_2^2}{n_2})^2} \\ &= \frac{(\frac{7828.715^2}{15} + \frac{5318.922^2}{20})^2}{\frac{1}{14}(\frac{7828.715^2}{15})^2 + \frac{1}{19}(\frac{5318.922^2}{20})^2} \\ &= 23.31274 \\ \therefore t^* &= t(23.31274) \\ &\approx 1.714 \end{aligned}$$

4. Conclusion: since $|t| = 1.639 < t^* = 1.714$, we **cannot reject the null hypothesis**.

(c) The true means: $\mu_1 = 2838.85 < \mu_2 = 6982.859$ which means the population mean of the Southern half *is* greater than that of the Northern half. The test in (b) rejected the hypothesis that $\mu_2 > \mu_1$ and **did not** detect the difference between them.
I think the other students will reach the same conclusion as n_1, n_2 are large enough for the samples to follow t-distribution. □

Problem 3. A poll is conducted to gather information about the proportion of voters p who will vote for a candidate in a primary election. Assume that the number of voters in the population is very large.

Solution. (a) $z^* = 1.96$ for 95% C.I. For an estimate p^*, the margin of error

$$
\begin{aligned}
1.96\sqrt{\frac{p^*(1-p^*)}{n}} &\leq 0.10 \\
3.8416\frac{p^*(1-p^*)}{n} &\leq 0.01 \\
n &\geq \frac{3.8416}{0.01}p^*(1-p^*) \\
&\geq \frac{3.8416}{0.01}(\frac{1}{2})^2
\end{aligned}
$$

since the above expression maximizes at $p^* = \frac{1}{2}$.

Hence $n \geq 96.04$ or $n = 97$ is the smallest sample size that will guarantee that the margin of error for this confidence interval will be no more than 0.10.

(b) The expected # of successes $n \times p = 97 \times 0.20 = 19.4 > 10$
The expected # of failures $= 97 \times 0.80 = 12.8 > 10$. Since both are more than 10, the **condition to use the large-sample interval is satisfied**.

□

Problem 4. Let p be the probability that the die comes up with the side labeled six. A six-sided die is rolled 100 times and the side labeled six comes up 27 times. Therefore $\hat{p} = X/n = 0.27$.

Solution. (a) At 99% confidence interval, $z^* = 2.575829$. The CI fpr p is

$$
\begin{aligned}
&= \hat{p} \pm z^*\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \\
&= 0.27 \pm 2.575829\sqrt{\frac{0.27(1-0.27)}{100}} \\
&= (0.1556436, 0.3843564)
\end{aligned}
$$

(b) At $\alpha = 0.1$,

- Hypothesis $H_0 : p = \frac{1}{6}$ vs. $H_a : p \neq \frac{1}{6}$

- Test statistic

$$
\begin{aligned}
z^* &= \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \\
&= \frac{0.27 - 0.167}{\sqrt{\frac{0.27(0.73)}{100}}} \\
&= 2.320032
\end{aligned}
$$

- $PValue = 2 \times P(z > 2.320032) = 2 \times (1 - P(z \leq 2.320032)) = 2 \times (1 - 0.9898304) = 0.02033916$
- Since $PValue = 0.02033916 < \alpha$ so **we reject** $H_0 : p = \frac{1}{6}$.

$\square$

1 Project details

1. Dataset: I plan to use the Stroke Prediction Dataset that I found on Kaggle for my project. It has 5110 records. Of of the 12 variables, some interesting ones are:
   - Gender
   - Ever Married (Yes/No)
   - Age
   - Residence Type (Urban/Suburban/Rural)
   - BMI Level
   - Had stroke (Yes/No)
2. We could excplore questions like:
   - Effect of BMI on having stroke.
   - Correlation between Residence Type and BMI.
   - Whether people in Suburban and Urban setting are more prone to Stroke than those in Rural.
3. Some summarization:

```
mean(stroke$heart_disease)
[1] 0.05401174
> mean(stroke$age)
[1] 43.22661
```

4. A histogram of ages in the given dataset: