Homework set 7 - Solutions

Math 3200 – Renato Feres

Preliminaries

The Student’s $t$ Test. A very common problem in statistics is to estimate the population mean of a probability distribution given a sample, the subject of sections 7.1 and 7.2 of the statistics textbook. As explained there, this relates directly to the so-called $t$ test. For large samples we may assume, due to the central limit theorem, that the sample mean is normally distributed; one often considers sample size $n > 30$ as being big enough in practice. For relatively small samples, in order to make inferences based on the $t$ test, the underlying population itself must be normally distributed.

We illustrate the $t$ test with an example in R. Let us first generate some data:

```r
x=rnorm(100, mean=5, sd=10).
```

Thus $x$ consists of 100 values drawn from a normal distribution with mean 5 and standard deviation 10. Suppose that the null hypothesis is $H_0 : \mu = 4$ and the alternative hypothesis is $H_1 : \mu \neq 4$. Let us apply to $x$ the R function `t.test`:

```r
> t.test(x,mu=4,conf.level=0.95,alternative="two.sided")
```

One Sample t-test

data: x
t = 2.4136, df = 99, p-value = 0.01764
alternative hypothesis: true mean is not equal to 4
95 percent confidence interval:
4.425743 8.360940
sample estimates:
mean of x
6.393341

There is a lot of information contained in this output; in order to interpret it, let us first review the underlying theory. We assume that $X_1, \ldots, X_n$ is a random sample from an $N(\mu, \sigma^2)$ distribution with unknown parameters $\mu$, $\sigma^2$. The pivotal random variable for making inferences on $\mu$ is

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

which has a $t$-distribution with $n - 1$ degrees of freedom. (In statistics, a pivotal quantity is a function of the random sample and of possibly unknown parameters whose probability distribution does not depend on the unknown parameters.)
Now select a confidence level $1 - \alpha$. In the above example, this was specified inside the \texttt{t.test} by assigning the value \texttt{conf.level=0.95}. This is the default value in R; it wasn’t really necessary to write it. Then specify the null and alternative hypothesis. This was done above by assigning \texttt{mu=4} and \texttt{alternative='two.sided'}. It isn’t necessary to write the latter either, as \texttt{'two.sided'} is also a default value. (The other possibilities are \texttt{'greater'} and \texttt{'less'} for $H_1: \mu > 4$ and $H_1: \mu < 4$, respectively.)

Recall that $1 - \alpha = P\left[-t_{n-1,\alpha/2} \leq T \leq t_{n-1,\alpha/2}\right]$. From this we derive the $100(1 - \alpha)\%$ confidence interval for $\mu$:

$$100(1 - \alpha)\% \text{ CI} = \left[\bar{x} - t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}\right]$$

where $\bar{x}$ is the sample mean and $s$ is the standard deviation. Finally, recall that the $P$-value in this (two-sided) case is the probability

$$P(|T_{n-1}| \geq |t|)$$

where $t$ is the sample value of $T$. That is, the probability under $H_0$ that $T$ takes on values at least “as extreme” as the one observed.

The output of \texttt{t.test} should now be self-explanatory: $t = 2.4136$, $df = 99$, $p$-value = 0.01764 are the $t$-statistic, degrees of freedom, and $P$-value. The confidence interval at confidence level 0.95 has endpoints 4.425743 \ and 8.360940 and the sample mean is 6.393341. For the same sample, an upper one-sided $t$-test at the 0.99 confidence level can be obtained thus:

```r
> t.test(x,mu=4,conf.level=0.99,alternative='greater')
```

\textbf{One Sample t-test}

data: x
t = 2.4136, df = 99, p-value = 0.008818
alternative hypothesis: true mean is greater than 4
99 percent confidence interval:
  4.048539   Inf
sample estimates:
  mean of x
6.393341

In both examples, the $P$-value is smaller than $\alpha$, so we are led to reject the null hypothesis at the respective significance levels.

\textbf{Checking for normality.} Before doing a $t$-test or other procedures that depend on a random variable being normal, it may be a good idea to determine whether the sample is normally distributed to a satisfactory degree. In examples 7.7 and 7.8 (pages 250 and 251 of the textbook) it is suggested the following procedure: look at the box plot and normal plot for confirmation that the distribution is roughly normal. If there are outliers, taken them out of the data set and see whether the normal plot improves (that is, becomes more straight).

Recall that, given sample data vector $x$, the box plot is obtained thus: \texttt{boxplot(x)} and the normal plot can be drawn as follows:

```r
qqnorm(x)
qqline(x)
```
Inferences on variance. The pivotal random variable for estimating population variance is chi-squared. If $S^2$ is the sample variance for a random sample $X_1, \ldots, X_n$ from an $N(\mu, \sigma^2)$ distribution, recall that
\[
\chi^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1},
\]
a chi-squared distribution with $n - 1$ degrees of freedom. Inferences on variance using the chi-squared distribution are strongly dependent on the assumption of normality.

Confidence intervals for $\sigma^2$. From the definition of the lower and upper critical points $\chi^2_{n-1,1-\alpha/2}$ and $\chi^2_{n-1,\alpha/2}$ we obtain (see page 255)
\[
100(1-\alpha)\% \ CI = \left[ \frac{(n-1)S^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1,\alpha/2}} \right],
\]
the confidence interval for $\sigma^2$ at the confidence level $1 - \alpha$. One-sided confidence intervals are similarly obtained. (See page 255.)

Recall that a critical point $\chi^2_{\nu,\alpha}$ can be obtain in R using the quartile function `qchisq` as follows:
\[
\chi^2_{\nu,\alpha} = qchisq(1-\alpha,\nu).
\]
For example,
\[
\chi^2_{9,0.5} = qchisq(0.95,9)=16.91898.
\]

Hypothesis tests on variance and $P$-value. This is explained on page 256 of the statistics textbook.

Problems

1. Exercise 7.13, page 264. A gasoline company tested 20 samples of gasoline produced during a day to check whether the day’s production meets the nominal octane rating of 87. The results were as follows (recall that the data files are provided at the link under the heading “Text” on the online syllabus):

<table>
<thead>
<tr>
<th>Octane</th>
<th>1</th>
<th>87.5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
<td>86.9</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>86.6</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>87.3</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>87.9</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>86.7</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>87.2</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>87.0</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>88.1</td>
</tr>
<tr>
<td></td>
<td>12</td>
<td>87.5</td>
</tr>
<tr>
<td></td>
<td>13</td>
<td>86.5</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>87.7</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>88.0</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>87.1</td>
</tr>
</tbody>
</table>
(a) Find a lower 95% confidence limit on the mean octane rating. Use this confidence limit to determine if the mean octane rating exceeds 87.

(b) Set up the hypotheses to show that the mean octane rating exceeds 87. Calculate the \( t \)-statistic and find bounds on its \( P \)-value. Is the result significant at \( \alpha = 0.005 \)? At \( \alpha = 0.001 \)?

**Solution:** Recall that the lower 95% CI for the mean, for \( n = 20 \), is

\[
\mu \geq \bar{x} - t_{19,0.05} \frac{s}{\sqrt{20}}.
\]

(a) We use R here and below simply as a calculator. The first step is to download the data into R. I will call the data vector (or table) by \( D \). The sample mean and standard deviations are as follows (Note: I first tried \( \text{mean}(D) \), which gave the correct value for the mean, with a message indicating that this command has been deprecated, and giving the preferred form, \( \text{sapply}(D,\text{mean}) \).)

\[
\begin{align*}
\text{colnames}(D) & = \text{'Octane'} \\
xbar & = \text{sapply}(D,\text{mean}) \\
s & = \text{sapply}(D,\text{sd}) \\
xbar & \\
\text{Octane} & \\
87.395 \\
s & \\
\text{Octane} & \\
0.5175601
\end{align*}
\]

The critical \( t \)-value can be obtained using the quantile function \( \text{qt} \). It is \( t_{19,0.05} = t_{\text{crit}} = \text{qt}(1-0.05,19) \). We obtain \( t_{\text{crit}} = 1.729133 \). Thus the lower 95% CI for \( \mu \) is

\[
\mu \geq \bar{x} - t_{19,0.05} \frac{s}{\sqrt{20}} = 87.195.
\]

Therefore, the mean octane rating exceeds 87 at the 0.95 confidence level.

(b) In this case, I will simply use the R function \( \text{t.test} \). Note that the null hypothesis is \( H_0 : \mu = 87 \) and the alternative hypothesis is \( H_1 : \mu > 87 \). Then

\[
\begin{align*}
\text{t.test}(D,\mu=87,\text{conf.level}=0.95,\text{alternative}=\text{'greater'})
\end{align*}
\]

One Sample \( t \)-test

\[
\begin{align*}
data: & \quad D \\
t & = 3.4131, \ df = 19, \ p-value = 0.001458
\end{align*}
\]
alternative hypothesis: true mean is greater than 87
95 percent confidence interval:
 87.19489    Inf
sample estimates:
mean of x
 87.395

This shows that the test statistic (the $t$-statistic in this case) is $t=3.4131$, the $P$-value is $p$-value=0.001458, meaning that

$$P(T_{19} > t) = 0.001458.$$ 

Therefore, the conclusion that $\mu > 87$ is valid at the significance level $\alpha = 0.005$, but not at the level $\alpha = 0.001$.

2. **Problem 7.15, page 265.** In response to student complaints and financial considerations, a high school decides to close its kitchen and contract a food service to provide school lunches. The previous year, when food was prepared in the high school kitchen, about 60% of the students purchased lunch on a daily basis. The daily proportions of students using the food service during the fourth month of the contract are given below.

<table>
<thead>
<tr>
<th>Proportions</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
</tbody>
</table>

Compared to the previous year, is there an increase in the average proportion of students purchasing lunches provided by the food service?

(a) The hypotheses are set up as $H_0 : \mu \leq 60$ vs. $H_1 : \mu > 60$. Explain the meaning of $\mu$, the parameter under test.

(b) Perform a test of the hypotheses in (a) using $\alpha = 0.01$. Interpret your results.
(c) The food service set a goal of attracting at least 70% of the students to purchase lunches. Set up the hypotheses to test whether the food service has met its goal.

(d) Calculate the $P$-value for the test of the hypotheses in (c). Is the result significant at $\alpha = 0.10$?

Solution: 

(a) The mean parameter $\mu$ is the true average proportion of students using the food provided by the food service.

(b) We wish to perform a test of the hypotheses at the level $\alpha = 0.01$. We reject $H_0$ at this level if the test statistic $t$ is greater than $t_{n-1,\alpha}$ and do not reject otherwise. First note that the critical point for $t$ is $t_{crit}=qt(1-0.01,19)=2.539483$. To find the test statistic we use `t.test`:

```r
> D=Ex7_15
> colnames(D)= 'Proportions'
> tcrit=qt(1-0.01,19)
[1] 2.539483
> t.test(D,mu=60,conf.level=0.99,alternative='greater')
```

One Sample t-test

```
data:  D
t = 5.7069, df = 19, p-value = 8.409e-06
alternative hypothesis: true mean is greater than 60
99 percent confidence interval:
  64.7176          Inf
sample estimates:
mean of x
  68.5
```

First observe that the $t$-statistic is $t=5.7069$, which is considerably bigger than the critical point $t_{crit}=2.5395$. This means that we can reject the null hypothesis in favor of $H_1: \mu \geq 60$ at level $\alpha = 0.01$.

(c) We wish to determine whether the food service has met the goal of attracting 70% of the students. In this case it makes sense to set $H_0: \mu \leq 70$ vs. $H_1: \mu > 70$. To fail to reject the null hypothesis is to fail in concluding that the goal was met.

(d) The $P$-value for the hypotheses of part (c) is the probability $P(T_{19} > t)$, where now the test statistic $t$ is calculated under the assumption $\mu = 70$.

```r
> t.test(D,mu=70,conf.level=0.10,alternative='greater')
```

One Sample t-test

```
data:  D
t = -1.0071, df = 19, p-value = 0.8367
alternative hypothesis: true mean is greater than 70
10 percent confidence interval:
  70.47757          Inf
```
The above gives a P-value of 0.8367. This is much bigger than $\alpha = 0.1$. So we cannot reject the null-hypothesis. We must conclude (for any reasonable significance level) that the goal has not been achieved.

3. **Problem 7.16, page 265.** A thermostat used in an electrical device is to be checked for the accuracy of its design setting of $200^\circ F$. Ten thermostats were tested to determine their actual settings, resulting in the following data:

<table>
<thead>
<tr>
<th>Temperatures</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Temperature</td>
<td>202.2</td>
<td>203.4</td>
<td>200.5</td>
<td>202.5</td>
<td>206.3</td>
<td>198.0</td>
<td>203.7</td>
<td>200.8</td>
<td>201.3</td>
<td>199.0</td>
</tr>
</tbody>
</table>

Perform the $t$-test to determine if the mean setting is different from $200^\circ F$. Use $\alpha = 0.05$.

**Solution:** Here I will simply use `t.test`. The null hypothesis is $H_0: \mu = 200$ and the alternative hypothesis is $H_1: \mu \neq 200$.

```r
> D=Ex7_16
> t.test(D,mu=200,conf.level=0.95,alternative='two.sided')
```

One Sample t-test

data:  D
t = 2.3223, df = 9, p-value = 0.04531
alternative hypothesis: true mean is not equal to 200
95 percent confidence interval:
    200.0459 203.4941
sample estimates:
mean of x
    201.77

The conclusion is that we may reject the null hypothesis at $\alpha = 0.05$ since the P-value is 0.04531 < $\alpha$. There is sufficient evidence at this significance level to conclude that the thermostat setting is different from 200 degrees.

4. **Problem 7.17, page 265.** Diabetic patients monitor their blood sugar levels with a home glucose monitor which analyses a drop of blood from a finger stick. Although the monitor gives precise results in a laboratory, the results
are too variable when it is used by patients. A new monitor is developed to improve the precision of the assay results under home use. Home testing on the new monitor is done by 25 persons using drops from a sample having a glucose concentration of 118 mg/dl. If $\sigma < 10$ mg/dl, then the precision of the new device under home use is better than the current monitor. The readings from 25 tests are as follows.

<table>
<thead>
<tr>
<th>glucose concentration</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>3</td>
</tr>
<tr>
<td>4</td>
</tr>
<tr>
<td>5</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>7</td>
</tr>
<tr>
<td>8</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>10</td>
</tr>
<tr>
<td>11</td>
</tr>
<tr>
<td>12</td>
</tr>
<tr>
<td>13</td>
</tr>
<tr>
<td>14</td>
</tr>
<tr>
<td>15</td>
</tr>
<tr>
<td>16</td>
</tr>
<tr>
<td>17</td>
</tr>
<tr>
<td>18</td>
</tr>
<tr>
<td>19</td>
</tr>
<tr>
<td>20</td>
</tr>
<tr>
<td>21</td>
</tr>
<tr>
<td>22</td>
</tr>
<tr>
<td>23</td>
</tr>
<tr>
<td>24</td>
</tr>
<tr>
<td>25</td>
</tr>
</tbody>
</table>

The sample mean is $\bar{x} = 118.5$ and the sample standard deviation is $s = 6.2$.

(a) Using R, check that the given values of $\bar{x}$ and $s$ are correct.

(b) Make a normal probability plot of these readings. Is it reasonable to assume that the data follow a normal distribution?

(c) Test $H_0 : \sigma \geq 10$ vs. $H_1 : \sigma < 10$ at the 0.10 level.

(d) Find an upper one-sided 90% confidence interval for $\sigma$. Use this interval to test the hypothesis in (b).

**Solution:** (a) The sample mean and sample standard deviation are:

```r
> colnames(D)=’glucose concentration’
> xbar=sapply(D,mean)
> s=sapply(D,sd)
> xbar
```
(b) A normal plot is obtained in R as follows:

```r
> D=Ex7_17
> x=D[,1]
> qqnorm(x)
> qqline(x)
> grid()
```

The result is shown in the below graph:
I would say that it is reasonable to assume that the data follows a normal distribution. For good measure, I also give the box plot (using `boxplot(D)`), even though it was not asked for. Notice that there are no indicated outliers and the plot seems reasonably symmetric. This is added confirmation that we are justified in assuming normality.

![Box Plot]

(c) We now wish to test $H_0 : \sigma \geq 10$ vs. $H_1 : \sigma < 10$ at the 0.10 level. The sample standard deviation is

\[ s = \text{sapply}(D, \text{sd}) = 6.192199. \]

The test statistic is

\[ \chi^2 = \frac{(n-1)s^2}{\sigma^2_0} = \frac{24 \times 6.1922^2}{10^2} = 9.2024. \]

The lower critical points is

\[ \chi^2_{n-1,1-\alpha} = qchisq(0.1, 24) = 15.6587. \]

Now, since $\chi^2 < \chi^2_{24,0.9}$ we have enough evidence to reject $H_0$ at the level $\alpha = 0.1$.

Although not asked, we also obtain the P-value. This is the probability that $\chi^2_{24} \leq \chi^2 = 9.02$. In R this can be found by `pchisq(9.02, 24)`, which gives the value 0.00245. This value is considerably less than 0.1, so the hypothesis $\sigma = 10$ is rejected at $\alpha = 0.1$.

(d) The upper one-sided confidence interval for $\alpha = 0.1$ is

\[ \sigma^2 \leq \frac{(n-1)s^2}{\chi_{n-1,1-\alpha}^2} = \frac{24 \times 6.19^2}{15.66} = 58.722. \]

Thus the upper one-sided interval for $\sigma$ is $\sigma \leq \sqrt{58.722} = 7.66$. Since this is less than $\sigma_0 = 10$, the conclusion is the same: reject $H_0$. 

10