Name Yiran Wang
NetID wang2559

Homework 1

STAT 601
September 28, 2019

Problem 1

Brief discussion over three common methods for estimating the distribution parameters:

(a) Method of Moments
In this method, moments from the sample are employed to estimate corresponding parameters from the population. The easiest example would be using the first origin moment of the sample to estimate the population mean while using the second central moment of the sample to estimate the population variance. Information about parameters is got from $E(X^k)$, that is:

$$\begin{cases} g_1(\theta_1,\cdots,\theta_k) = E(X) \\ \quad\quad\quad\cdots\cdots \\ g_k(\theta_1,\cdots,\theta_k) = E(X^k) \end{cases}$$

The solutions for $\theta_i, i = 1,\cdots,k$ are:

$$\theta_i = h_i(E(X),\cdots,E(X^k))$$

Moment estimator:

$$\widehat{\theta_i} =: h_i(\overline{X},\cdots,\overline{X^k})$$

- Pros: Method of moments is fairly simple and is easy to compute. It is asymptotically normal. Distribution of population is not required. It is a consistent estimator.
- Cons: This method is not necessarily the most efficient estimate. In some cases with small samples, the estimates given by method of moments may be outside of the valid range. The estimates may not be unique.

(b) Least Squares
For any line $y = c + dx$, the residual sum of squares(RSS) is defined as

$$RSS = \sum_{i=1}^{n}(y_i - (c + dx_i))^2$$

The least squares estimates of α and β are defined to be those values a and b such that the line $a + bx$ minimizes RSS. That is, the least square estimates, a and b, satisfy

$$\min_{c,d}\sum(y_i - (c + dx_i))^2 = \sum_{i=1}^{n}(y - (a + bx_i))^2$$

- Pros: This method is fairly easy for estimate unknown data while minimize the residual sum of squares.

- Cons: This method considers only observational errors in the dependent variable. Besides, if there are many outliers in observations, the estimates by the least squares method will have big error.

(c) Maximum Likelihood
Consider a random sample $X_1, X_2, ..., X_n \sim f(x, \theta), \theta \in \Omega$. The joint p.d.f. of $X_1, X_2, \cdot\cdot\cdot, X_n$ is

$$f(x_1, \theta)f(x_2, \theta) \cdots f(x_n, \theta)$$

This joint p.d.f. can be regarded as a function of θ, and it is called the likelihood function L of the random sample. Write for $\theta \in \Omega$,

$$L(\theta; x_1, x_2, \cdot\cdot\cdot, x_n) = f(x_1, \theta)f(x_2, \theta) \cdots f(x_n, \theta)$$

Suppose that $u(x_1, x_2, \cdot\cdot\cdot, x_n)$ is a nontrivial function such that, when θ is replaced by $u(x_1, x_2, \cdot\cdot\cdot, x_n)$, the likelihood function L is maximized:

$$L(u(x_1, x_2, \cdot\cdot\cdot, x_n); x_1, x_2, \cdot\cdot\cdot, x_n) = \max_{\theta \in \Omega} L(\theta; x_1, x_2, \cdot\cdot\cdot, x_n)$$

Then the statistic $u(X_1, X_2, \cdot\cdot\cdot, X_n)$ will be called a maximum likelihood estimator (MLE) of θ and will be denoted by the symbol $\widehat{\theta} = u(X_1, X_2, \cdot\cdot\cdot, X_n)$.

- Pros: It is asymptotically unbiased, consistent, normally distributed, and efficient. Estimator given MLE minimized the variance among all other unbiased estimators.
- Cons: It can be highly biased for small samples. Sometimes, MLE has no closed-form. Assumptions for probability distributions function of every variable are required. It's often hard to calculate the result of MLE.

(d) The MLE of the mean and the variance of the univariate Gaussian distribution:
Let $X_1, X_2, \cdot\cdot\cdot, X_n$ iid.$\sim N(\mu, \sigma^2), -\infty < \theta_1 < \infty, -\infty < \theta_2 < \infty$. Since the p.d.f. of $N(\mu, \sigma^2)$ is:

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp[-\frac{(x-\mu)^2}{2\sigma^2}]$$

Thus, the likelihood function based on the observations is

$$L(\mu, \sigma^2) = \frac{1}{(\sqrt{2\pi}\sigma)^n} exp[-\sum_{j=1}^{n} \frac{(x_j-\mu)^2}{2\sigma^2}]$$

Maximize the logarithm of the likelihood function by differentiation equations:

$$lnL(\mu, \sigma^2; x_1, \cdot\cdot\cdot, x_n) = -\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{2\sigma^2} - \frac{nln(2\pi\sigma^2)}{2}$$

$$\frac{\partial lnL}{\partial \mu} = \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0$$

$$\frac{\partial lnL}{\partial \sigma^2} = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^4} - \frac{n}{2\sigma^2} = 0$$

By calculation, the solutions for μ and σ^2 are:

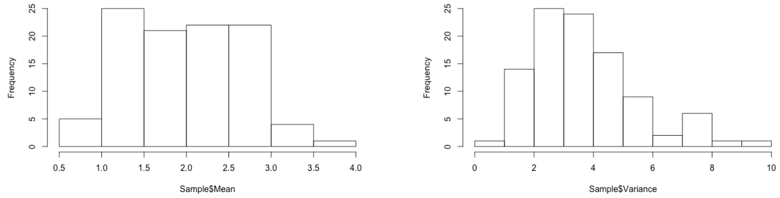
$$\mu = \frac{\sum_{i=1}^{n} x_i}{n} =: \overline{x}$$

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} =: s^2$$

Thus, the MLEs of μ and σ^2 are, respectively, the mean and the variance of sample, namely: $\overline{\mu} = \overline{X}$; $\overline{\sigma}^2 = S^2$.

Problem 2

(a) Histograms of sample mean and sample variance for $N(2,4)(n = 10)$

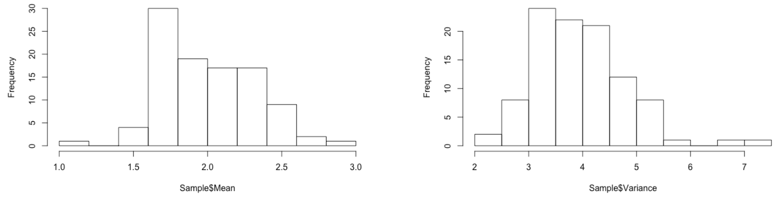


(a) Mean

(b) Variance

Figure 1: Histograms of sample mean and sample variance for $N(2,4)(n = 10)$

(b) Histograms of sample mean and sample variance for $N(2,4)(n = 40)$

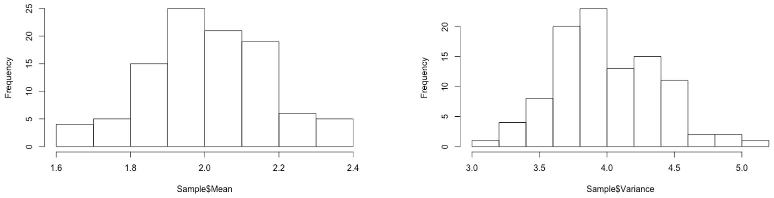


(a) Mean

(b) Variance

Figure 2: Histograms of sample mean and sample variance for $N(2,4)(n = 40)$

(c) Histograms of sample mean and sample variance for $N(2,4)(n = 160)$



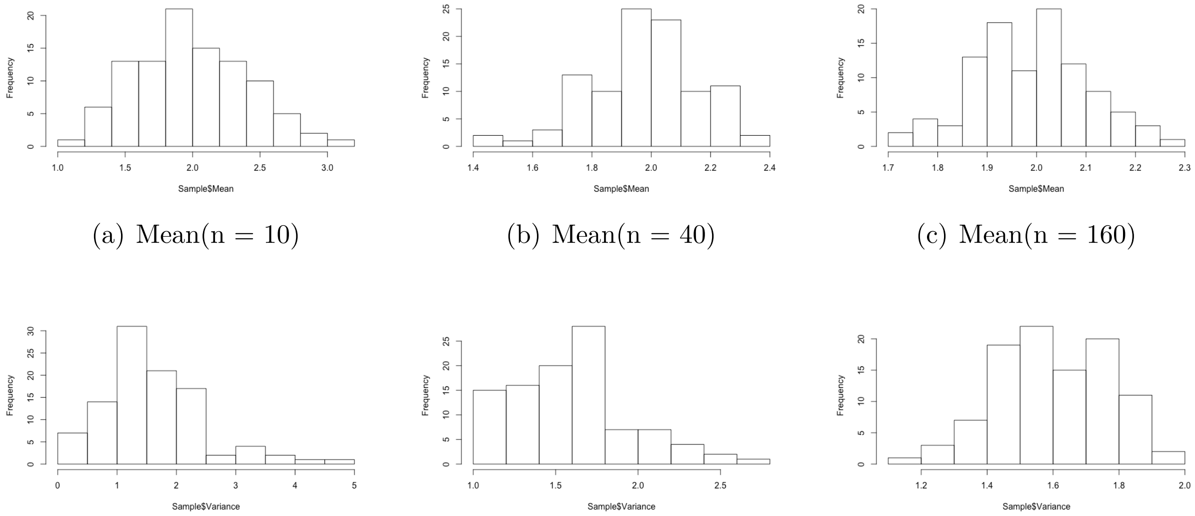
(a) Mean

(b) Variance

Figure 3: Histograms of Sample Mean and Variance for $N(2,4)(n = 160)$

(d) Observations made from (a)-(c):
As the sample size increases, the histograms of these random sample approach the values of parameters from the given distribution $N(2,4)$. The means of histograms have a tendency towards 2, while the vairances approach 4.

(e) Replace $N(2,4)$ with $B(10,0.2)$:
As the sample size increases, the histograms of these random sample approach the values of parameters from the given distribution $B(10,0.2)$. According to the properties of binomial distribution, if $X \sim B(n,p)$, in this case,$B(10,0.2)$, then $E(X) = np = 2$, $Var(X) = np(1-p) = 1.6$, which corresponds with the results below.



(a) Mean(n = 10)

(b) Mean(n = 40)

(c) Mean(n = 160)

(d) Variance(n = 10)

(e) Variance(n = 40)

(f) Variance(n = 160)

Figure 4: Histograms of Sample Mean and Variance for $B(10,0.2)$

(f) Find $E(\overline{Y})$, $Var(\overline{Y})$, and $E(S^2)$.
Because $Y_1, \cdots, Y_n$ are i.i.d. samples from $N(\mu, \sigma)$ and according to the properties of mean and variance:

$$E(\overline{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}(\mu + \mu + \cdots + \mu) = \mu$$

$$Var(\overline{Y}) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2}n\sigma = \frac{\sigma}{n}$$

$$\begin{aligned}
E(S^2) &= E(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2) \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}[(Y_i - \mu) - (\overline{Y} - \mu)]^2\right\} \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2\sum_{i=1}^{n}(Y_i - \mu)(\overline{Y} - \mu) + n(\overline{Y} - \mu)^2] \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n}(Y_i - \mu)^2 - 2n(\overline{Y} - \mu)^2 + n(\overline{Y} - \mu)^2] \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n} E(Y_i - \mu)^2 - nE(\overline{Y} - \mu)^2\right\} \\
&= \frac{1}{n-1}[n\sigma - n(\frac{\sigma}{n})] \\
&= \sigma
\end{aligned}$$

From equations above, for $N(2,4)$, we have $E(\overline{Y}) = 2$, $Var(\overline{Y}) = 4/n$, and $E(S^2) = 4$. According to the simulation results, we have:

Sample Size	E(Y)	Var(Y)	E(S)
N	2	4/N	4
10	1.9047657	0.4060186	4.0445334
40	1.9918165	0.1011323	3.9206122
160	1.97246826	0.03016786	3.95059983

Table 1: Comparison between simulated and theoritical values for $N(2,4)$

By comparison, we can see that for $N(2,4)$, as the sample size becomes larger, the mean of $\overline{Y}$ converges to the population mean and the variance of $\overline{Y}$ converges to the population variance divided by n.

The difference between $Var(\overline{Y})$ and $Var(Y)$:
$Var(\overline{Y})$ refers to the variance of the sample mean, while $Var(Y)$ is the population variance. The value of $Var(\overline{Y})$ equals to the value of $Var(Y)$ divided by n.

(g) Find $E(\overline{Y})$, $Var(\overline{Y})$, and $E(S^2)$.
Because $Y_1, \cdot\cdot\cdot, Y_n$ are i.i.d. samples from $B(m, \pi)$ and according to the properties of mean and variance:

$$E(\overline{Y}) = E(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n}(m\pi + m\pi + \cdots + m\pi) = m\pi$$

$$Var(\overline{Y}) = Var(\frac{1}{n}\sum_{i=1}^{n} Y_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(Y_i) = \frac{1}{n^2} nm\pi(1-\pi) = \frac{m\pi(1-\pi)}{n}$$

$$\begin{aligned}
E(S^2) &= E(\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2) \\
&= \frac{1}{n-1}E\left\{\sum_{i=1}^{n}[(Y_i - m\pi) - (\overline{Y} - m\pi)]^2\right\} \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n}(Y_i - m\pi)^2 - 2\sum_{i=1}^{n}(Y_i - m\pi)(\overline{Y} - m\pi) + n(\overline{Y} - m\pi)^2] \\
&= \frac{1}{n-1}E[\sum_{i=1}^{n}(Y_i - m\pi)^2 - 2n(\overline{Y} - m\pi)^2 + n(\overline{Y} - m\pi)^2] \\
&= \frac{1}{n-1}\left\{\sum_{i=1}^{n}E(Y_i - m\pi)^2 - nE(\overline{Y} - m\pi)^2\right\} \\
&= \frac{1}{n-1}[nm\pi(1-\pi) - n(\frac{m\pi(1-\pi)}{n})] \\
&= m\pi(1-\pi)
\end{aligned}$$

From equations above, for $B(10, 0.2)$, we have $E(\overline{Y}) = 2$, $Var(\overline{Y}) = 1.6/n$, and $E(S^2) = 1.6$. According to the simulation results, we have:

Sample Size	$E(\overline{Y})$	$Var(\overline{Y})$	$E(S^2)$
N	2	1.6/N	1.6
10	2.0240000	0.1485091	1.5782222
40	1.98400000	0.05500657	1.55978205
160	2.00487500	0.01038635	1.58208176

Table 2: Comparison between simulated and theoritical values for $B(10, 0.2)$

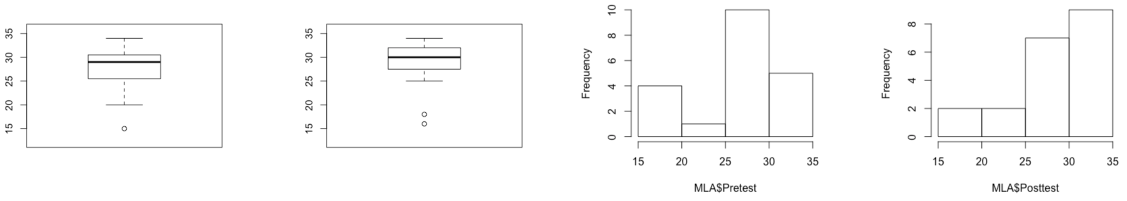
Similarly, we can conclude that as the sample size increases, the value calculated from the samples converges to the theoritical value.
The difference between $Var(\overline{Y})$ and $Var(Y)$:
$Var(\overline{Y})$ refers to the variance of the average of Y, while $Var(Y)$ is the population variance. The value of $Var(\overline{Y})$ equals to the value of $Var(Y)$ divided by n. Although $\overline{Y}$ and Y share mutual mean value, but their stretchabilities are different.

Problem 3

(a) To show that the training improves listening skills, we denote:
Y_{1i}: Random variable of scores on the MLA listening pretest for $i = 1, \cdots, 20$.
Y_{2i}: Random variable of scores on the MLA listening posttest for $i = 1, \cdots, 20$.
$\mu_1 = E(Y_{1i})$: Population mean MLA scores in the pretest.
$\mu_2 = E(Y_{2i})$: Population mean MLA scores in the posttest.
$D_i = Y_{1i} - Y_{2i}$: MLA score difference of the i^th executive between the pretest and the posttest.
$\mu_D = E(D_i) = \mu_1 - \mu_2$: Population mean of MLA score difference between the pretest and the posttest.
Equivalent to testing $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 < \mu_2$, we test

$$H_0 : \mu_D = 0 \quad vs. \quad H_A : \mu_D < 0$$

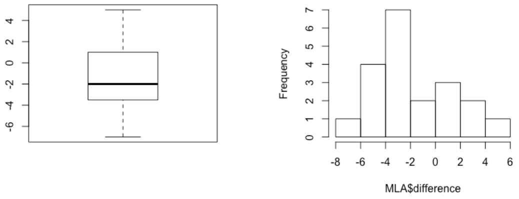
(b) From barplots, it is clear that the posttest scores are left-skewed. Based on boxplots, we can see that there is 1 outlier in the pretest data (Subject = 8) and 2 outliers in the posttest data (Subject = 5, 8 respectively). However, when we check the two subjects individually, we can see that their scores are not invalid because the difference between their pretest and posttest is reasonable and plausible. Thus, we should not drop them from the data set.



(a) Boxplot_Pretest (b) Boxplot_Posttest (c) Histogram_Pretest (d) Histogram_Posttest

Figure 5: Boxplots and Histograms of Raw MLA Scores

Since we will use Paired T test, we should actually make assumptions on the difference of pretest and posttest. Accordingly, there are no outliers in score difference, while this data is slightly right-skewed. Since Paired T Test is robust against non-normality, and pairs are independent of each other pair (the score for each executive is independent), this test is valid.



(a) Boxplot (b) Histogram

Figure 6: Boxplot and Histogram of MLA Score Difference

(c) A statistic is the sample mean MLA score difference $\overline{D}$ based on an i.i.d. sample of size $n = 20$. Assume that the $H_0 : \mu_D = 0$ holds, and assume that $D_i \sim N(0, \sigma^2)$. Thus, under H_0,

$$T = \frac{\overline{D} - \mu_D^0}{\frac{S_D}{\sqrt{n}}} \sim T_{n-1}$$

where T_{n-1} is a T distribution with $n-1$ degrees of freedom. From the summary statistics, we have $n = 20$, $\overline{d} = -1.45$, and $s_D = 3.203206$. Thus, the standard error is

$$\frac{s_d}{\sqrt{n}} = 0.7162586$$

The observed test statistic is

$$t = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = -2.024409$$

Compute the p-value from Table B, and we have:

$$P(T_{19} \geqslant -2.024409) = 0.02861$$

We then reject H_0 at 5% level.

(d) Suppose $D_1, D_2, \cdots, D_n$ is an i.i.d. sample from $N(\mu_D, \sigma_D^2)$ and σ_D^2 is unknown. Let $t_{n-1,\frac{\alpha}{2}}$ denote the t critical value such that

$$P(-t_{n-1,\frac{\alpha}{2}} \leqslant T_{n-1} \leqslant t_{n-1,\frac{\alpha}{2}}) = 1 - \alpha$$

Then we have a $(1-\alpha) = 90\%$ confidence interval for μ_D as

$$\mu_D \in [\overline{d} - t_{n-1,\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}, \overline{d} + t_{n-1,\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}},]$$

In this question, a 90% CI for μ_D is

$$-1.45 - 1.729 \times \frac{3.203206}{\sqrt{20}} \leqslant \mu_D \leqslant -1.45 + 1.729 \times \frac{3.203206}{\sqrt{20}}$$

which is [-2.688411,-0.211589].

(e) Since $D_i \sim N(2,1)$, that is, the variance is known. For $i = 1, \cdot \cdot \cdot, n$. $\alpha = 0.05$, $1 - \beta = 0.80$,

$$T(n) = \frac{\overline{d} - \mu_0}{\frac{s_d}{\sqrt{n}}} = \frac{-1.45 - 0}{\frac{3.203206}{\sqrt{n}}} = -0.4526715\sqrt{n} \sim N(0,1)$$

Because $Z_{\frac{\alpha}{2}} = 1.96$, $u_A - u_0 >> 0$,then we have

$$1 - \beta = 0.2 \geqslant \phi(\frac{\mu_A - \mu_0}{\sigma/n} - z_{\alpha/2})$$

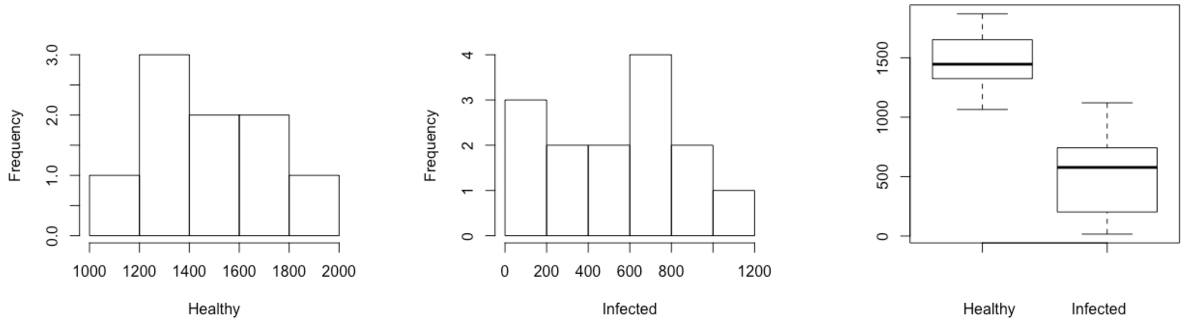
By calculation in R, then we have

$$n = (\frac{Z_{\alpha/2} + Z_\beta}{(\mu_A - \mu_0)/\sigma})^2$$

Thus, the minimum sample size n needed is 2.

Problem 4

(a) Tentative conclusions:
According to the summary of some statistics and plots below, minimum, 1^{st} Quantile, median, mean, 3^{rd} Quantile and maximum of the healthy seedlings are all apparently greater than those of the infected, but the standard error contradicts. Thus, we tend to believe that there is a severe negative effect from the virus infection that hinders the growth of these seedlings significantly. Besides, the negative effect varies greatly among these seedlings.



(a) Hist of Healthy

(b) Hist of Infected

(c) Boxplot

Figure 7: Boxplot and Histogram of Stem Volume

By calculation in R, we have:

Group	N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Healthy	9	1065	1325	1446	1480	1652	1870	248.98
Infected	14	16.0	235.2	577.5	549.4	741.8	1121.0	343.46

(b) Considering that the given data are unpaired, we choose independent two sample T test with equal variance. Notations are as below:
Y_{1i}: Random variable of stem volume of the ith 2-year-old seedlings propagated from healthy buds, $i = 1, \cdots, 9$.
Y_{2i}: Random variable of stem volume of the ith 2-year-old seedlings propagated from virus-infected buds, $i = 1, \cdots, 14$.
$\mu_1 = E(Y_{1i})$: Population mean stem volume in the healthy group.
$\mu_2 = E(Y_{2i})$: Population mean stem volume in the infected group.
Our goal is to test

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 > \mu_2$$

Consider the difference in mean $\overline{Y}_1 - \overline{Y}_2$. Under the H_0, the test statistic follows t-distribution with $df = n_1 + n_2 - 2$.

$$t = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{\sqrt{S_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} \sim T_{n_1+n_2-2}$$

From the given data, we have

$$\overline{y}_1 = 1480,\ s_1^2 = 61993.5,\ n_1 = 9 \quad and \quad \overline{y}_2 = 549.4,\ s_2^2 = 117963.8,\ n_2 = 14.$$

The pooled sample variance is

$$S_p^2 = \frac{(9-1)\times 61993.5 + (14-1)\times 117963.8}{9+14-2} = 96641.78$$

Thus, the observed test statistic is:

$$t = \frac{1480 - 549.4 - 0}{\sqrt{96641.78(\frac{1}{9}+\frac{1}{14})}} = 7.006298$$

The degrees of freedom are: $df = 9 + 14 - 2 = 21$.
The p-value is $P(T_{21} \geqslant 7.006298)$, which is 3.223e-07, less than 0.001.
The conclusion is: Reject H_0 at 5% level. There is very strong evidence that the mean stem volume from infected buds is smaller than that from healthy buds.

(c) Sicne under the H_0, the test statistic follows t-distribution with $df = n_1 + n_2 - 2$, we then have a $(1-\alpha)$ CI for $\mu_1 - \mu_2$ is

$$\overline{Y}_1 - \overline{Y}_2 \ \pm t_{n_1+n_2-2,\alpha} \times \sqrt{s_p^2(\frac{1}{n_1}+\frac{1}{n_2})}$$

where $t_{n_1+n_2-2,\alpha}$ is the critical value for which $P(|T_{n_1+n_2-2}| \geq t_{n_1+n_2-2,\alpha}) = \alpha$.
Taken the real value into account, we have $t_{21,0.05} = 1.721$, $S_p^2 = 96641.78$. Thus, a 95% confidence interval for the difference of the mean stem volume of 2-year-old seedlings between the two groups is:

$$\mu_D \in [701.9895, 1800.847]$$

(d) Considering that the given data are unpaired, we choose independent two sample T test with equal variance. Notations are the same as (b), but in this case, our goal is to test

$$H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$$

Similarly,from the given data, we have

$$\bar{y}_1 = 1480,\ s_1^2 = 61993.5,\ n_1 = 9 \quad and \quad \bar{y}_2 = 549.4,\ s_2^2 = 117963.8,\ n_2 = 14.$$

The pooled sample variance is

$$S_p^2 = \frac{(9-1)\times 61993.5 + (14-1)\times 117963.8}{9+14-2} = 96641.78$$

Thus, the observed test statistic is:

$$t = \frac{1480 - 549.4 - 0}{\sqrt{96641.78(\frac{1}{9}+\frac{1}{14})}} = 7.006298$$

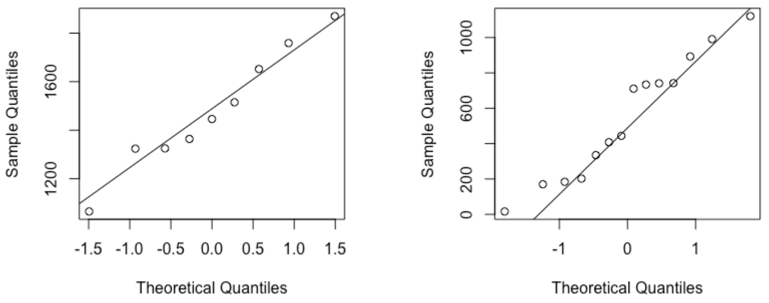
The degrees of freedom are: $df = 9 + 14 - 2 = 21$.
The p-value is $2 \times P(T_{21} \geqslant 7.006298)$, which is 3.223e-07, less than 0.001.
The conclusion is: Reject H_0 at 5% level. There is very strong evidence that the mean stem volume from healthy buds and infected buds are different.

(e) Assumptions made for (b), (c) and (d) are as below:

1) The healthy sample $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an i.i.d. sample of size n_1 from $N(\mu_1, \sigma_1^2)$.
2) The infected sample $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an i.i.d. sample of size n_2 from $N(\mu_2, \sigma_2^2)$.
3) The two samples Y_{1i} and Y_{2i} are independent.
4) The (unknown) variances are the same $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

Assess these assumptions:

1) Independence:
Consider the study or experimental design, the independence within and between the two samples are reasonable.
2) Normality:
In order to assess the normality, we choose QQ plot as a more reliable tool. From the plot below, since the QQ line of the infected group shows a transparent departure from the 45° line, we have that the infected group is not necessarily normal distributed.



(a) Healthy

(b) Infected

Figure 8: QQ Plot of Stem Volume From 2 groups

3) Equal variance:
Because we employed the unpaired independent T test with equal variances, thus we ought to assess their homoscedasticity first. By running codes on R, we have the p-value of $Levene'stest$ as 0.1112. Thus, $H_0 : \sigma_1^2 = \sigma_2^2$ is not rejected at 5%

level.
Similarly, from Figure 7, the stretchnesses of two samples are not significantly different from each other. Thus, we accept the equal variance assumption.

In brief, the independence and equal variance assumptions are reasonable for both group, but the normality for the infected group is not necessarily reasonable.
In terms of remedial measures, we could utilize log transformation to make the transformed data align better with the assumptions.

(f) Perform a Welch's T test:
Notations:
Y_{1i}: Random variable of stem volume of the ith 2-year-old seedlings propagated from healthy buds, $i = 1, \cdots, 9$.
Y_{2i}: Random variable of stem volume of the ith 2-year-old seedlings propagated from virus-infected buds, $i = 1, \cdots, 14$.
$\mu_1 = E(Y_{1i})$: Population mean stem volume in the healthy group.
$\mu_2 = E(Y_{2i})$: Population mean stem volume in the infected group.
To test $H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$, we test:

$$H_0 : \mu_1 - \mu_2 = 0 \quad vs. \quad H_A : \mu_1 - \mu_2 \neq 0$$

Assumptions:

1) The healthy sample $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an i.i.d. sample of size n_1 from $N(\mu_2, \sigma_2^2)$.
2) The infected sample $Y_{21}, Y_{22}, ..., Y_{2n_2}$ is an i.i.d. sample of size n_2 from $N(\mu_2, \sigma_2^2)$.
3) The two samples Y_{1i} and Y_{2i} are independent.
4) The (unknown) variances are not the same $\sigma_1^2 \neq \sigma_2^2$.

Still, we use a T-type test statistic:

$$T = \frac{\overline{Y}_1 - \overline{Y}_2 - 0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

Use T_{adf} to approximate the distribution above:

$$adf \approx \frac{(r_1 + r_2)^2}{\frac{r_1^2}{n_1 - 1} + \frac{r_2^2}{n_2 - 1}}$$

where $r_1 = \frac{s_1^2}{n_1}$ and $r_2 = \frac{s_2^2}{n_2}$.
We then perform Levene's test, and the result is:

$$f = 2.7656 \ on \ df = 21; \ p - value = 0.1112$$

We then perform the Welch's T test since the H_0 of $Levene'stest$ that two samples have equal variance is rejected on 5% level:

$$t = 7.5197 \ on \ df = 20.586, \ p - value = 2.487e - 07$$

Thus, the conclusion is reject H_0 at 5% level. There is very strong evidence that the mean of log stem volumn from healthy and infected buds are different. A 95% CI is [672.9032, 1188.2396].

(g) Perform a suitable randomization test as below:
Our goal is to test: $H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$.
Rearrange the data, add a column of type H(Healthy) and I(Infected). Calculate the sample difference: $\overline{y}_1 - \overline{y}_2 = 1480 - 549.4 = 930.6$.
Under $H_0 : \mu_1 = \mu_2$, randomly assign type H and I.
Sample vector of size n of 2 types with the replacement, and repeat 10,000 times, count the number of times that the sample mean difference is as large as or larger than the observed sample mean difference 930.6, and compute a p-value as

$$2 \times \frac{0}{10000} = 0$$

Conclusion: There is strong evidence against the H_0 that the mean stem volumn is the same for the two bud types. Reject the H_0 at the $\alpha = 0.05$ level.

(h) Because the given bud data is not paired, we choose Wilcoxon Test.
H_0: The two populations have the same distribution.
H_A: The two populations have the same shape but different locations.
Firstly, we merge the two samples into one combined sample. Then we sort the observations in the combined sample in ascending order and assign each observation a rank. Add the ranks corresponding to observations from the healthy group denoted as R_1. Observed rank sum is 170. We use R_1 as a test statistic. Based on the observed r_1, we then conduct a randomization.
Under $H_0 : \mu_1 = \mu_2$, randomly assign type H and I. Sample vector of size n of 2 types with the replacement, and repeat 10,000 times, count the number of times that the sample R_1 is as large as or larger than the observed sample $R_1 = 170$, and compute a p-value as

$$2 \times \frac{0}{10000} = 0$$

From R, p-value = 0.
Similarly, there is strong evidence against the H_0 that the two populations have the same distribution. Thus, we reject the H_0 at $\alpha = 0.05$ level.

(i) From (d), the result from two sample T test is p-value = 3.223e-07, we then reject $H_0 : \mu_1 = \mu_2$. However, the assumptions for this test are not all necessarily satisfied.

From (f), the result from Welch's T test is p-value = 2.487e−07, we then reject $H_0 : \mu_1 = \mu_2$.
From (g), the empirical p-value = 0, we then reject $H_0 : \mu_1 = \mu_2$.
From (h), the empirical p-value = 0, we then reject H_0 that the two populations have the same distribution.
Conclusively, we can say that the mean stem volumn for healthy and infected buds are significantly different.

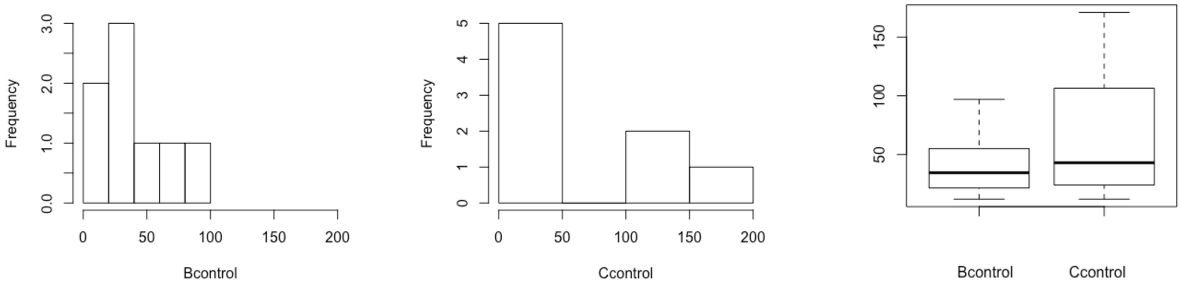
(j) Since the value of d is not given, according to thresholds for effect size from Cohen, we set $d = 0.8$, then we calculate the paired n_1 and n_2 in R. Results for minimum size of n_1 and n_2 are listed as below:

n_1	13	14	15	16	17	18	19	20	21	22	23	24	25	26
n_2	250	113	77	60	50	43	39	35	33	31	29	28	27	26

In order to make the unpaired two sample T test more robust against differences in variance, it would be better if we set the two sample sizes n_1 and n_2 as approximately equal.

Problem 5

(a) Tentative conclusions:
According to the summary of some statistics and plots below, minimum, 1^{st} Quantile, median, mean, 3^{rd} Quantile and maximum of numbers of moths captured from biological controlled plot are all apparently smaller than those from chemical controlled plot, but the standard error contradicts. Thus, we tend to believe that there is an effective reduction of the damage from moths under the biological control. Besides, the effectiveness is quite robust considering the reduced standard deviation.



(a) Hist of Biological Control (b) Hist of Chemical Control (c) Boxplot

Figure 9: Boxplot and Histogram of Moths under Biological and Chemical Control

By calculation in R, we have:

Group	N	Min	1st Qu.	Median	Mean	3rd Qu.	Max	SD
Biological control	8	12.00	25.25	34.5	41.38	48.50	97	28.48
Chemical control	8	12.00	27.00	43.00	66.25	106.25	171.00	55.92

(b) Considering that the given data are paired, we choose paired two sample T test. Notations are as below:
Y_{1i}: Random variable of moths number of the ith plot treated with biological control, $i = 1, \cdots, 8$.
Y_{2i}: Random variable of moths number of the ith plot treated with chemical control, $i = 1, \cdots, 8$.
$\mu_1 = E(Y_{1i})$: Population mean number of moths treated with biological control.
$\mu_2 = E(Y_{2i})$: Population mean number of moths treated with chemical control. $D_i = Y_{1i} - Y_{2i}$: Moths number difference of the i^th plot between the biological and the chemical control.
$\mu_D = E(D_i) = \mu_1 - \mu_2$: Population mean of moths number difference between biological and the chemical control.
Equivalent to testing $H_0 : \mu_1 = \mu_2$ vs. $H_A : \mu_1 \neq \mu_2$, we test

$$H_0 : \mu_D = 0 \quad vs. \quad H_A : \mu_D \neq 0$$

A statistic is the sample mean of moths number difference $\overline{D}$ based on an i.i.d. sample of size $n = 8$. Assume that the $H_0 : \mu_D = 0$ holds, and assume that $D_i \sim N(0, \sigma^2)$. Thus, under H_0,

$$T = \frac{\overline{D} - 0}{\frac{S_D}{\sqrt{n}}} \sim T_{n-1}$$

where T_{n-1} is a T distribution with $n-1$ degrees of freedom.
From the summary statistics, we have $n=8$, $\overline{d}=-24.875$, and $s_D=40.18$. Thus, the standard error is

$$\frac{s_d}{\sqrt{n}}=14.20568$$

The observed test statistic is

$$t=\frac{\overline{d}-0}{\frac{s_d}{\sqrt{n}}}=-1.75106$$

Compute the p-value from Table B, and we have:

$$2\times P(T_8\leqslant -1.75106)=0.1234$$

We then do not reject H_0 at 5% level.

(c) Suppose $D_1, D_2, \cdots, D_n$ is an i.i.d. sample from $N(\mu_D, \sigma_D^2)$ and σ_D^2 is unknown. Let $t_{n-1,\frac{\alpha}{2}}$ denote the t critical value such that

$$P(-t_{n-1,\frac{\alpha}{2}}\leqslant T_{n-1}\leqslant t_{n-1,\frac{\alpha}{2}})=1-\alpha$$

Then we have a $(1-\alpha)=95\%$ confidence interval for μ_D as

$$\mu_D\in[\overline{d}-t_{n-1,\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}},\overline{d}+t_{n-1,\frac{\alpha}{2}}\frac{S_D}{\sqrt{n}},]$$

In this question, a 95% CI for μ_D is

$$-24.875-2.365\times\frac{40.17973}{\sqrt{8}}\leqslant\mu_D\leqslant -24.875+2.365\times\frac{40.17973}{\sqrt{8}}$$

which is [-58.47143, 8.721433].

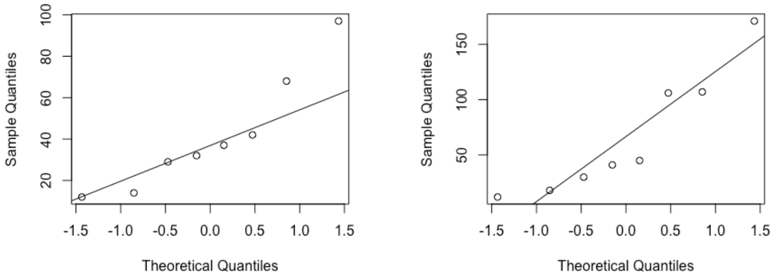
(d) Assumptions made for (b), (c) are as below:

1) The biological controlled sample $Y_{11}, Y_{12}, ..., Y_{1n_1}$ is an i.i.d. sample of size n from $N(\mu_1, \sigma_1^2)$.
2) The chemical controlled sample $Y_{21}, Y_{22}, ..., Y_{2n_1}$ is an i.i.d. sample of size n from $N(\mu_2, \sigma_2^2)$.
3) The two samples Y_{1i} and Y_{2i} are independent.

Assess these assumptions:

1) Independence:
Considering the study or experimental design, within each plot, subplots are randomly assigned to two types of control. Thus, the independence within and between the two samples are reasonable.

2) Normality:
In order to assess the normality, we choose QQ plot as a more reliable tool. From the plot below, since the QQ lines of two groups show a transparent departure from the 45° line, we say that these two groups are not necessarily normal distributed.

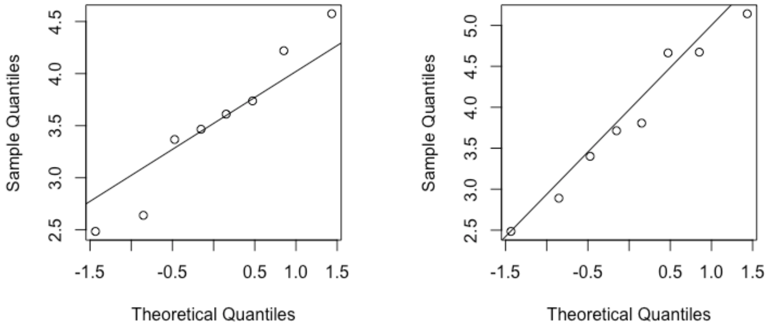


(a) Biological controlled (b) Chemical controlled

Figure 10: QQ Plots of Moth number From 2 Types of Control

In terms of remedial measures, we could utilize log transformation to make the transformed data align better with the normality assumption.

(e) Since the raw data are not necessarily normal distributed, we apply a log-transformation so that the transformed data can align better with the assumptions.
After transformation, new QQ plots are as below. We can see that after transformation, the data align better.



(a) Biological controlled (b) Chemical controlled

Figure 11: Transformed QQ Plots of Moth number From 2 Types of Control

Thus, we still use paired T test. From the summary statistics, we have $n = 8$, $\overline{log(d)} = -0.334355$, and $s_D = 0.4788016$. Thus, the standard error is

$$\frac{s_d}{\sqrt{n}} = 0.1692819$$

The observed test statistic is

$$t = \frac{\overline{d} - 0}{\frac{s_d}{\sqrt{n}}} = -1.9751$$

Compute the p-value from Table B, and we have:

$$2 \times P(T_8 \leqslant -1.9751) = 0.08882$$

We then still do not reject H_0 at 5% level.

(f) Our goal is to test: $H_0 : \mu_1 = \mu_2 \quad vs. \quad H_A : \mu_1 \neq \mu_2$.
Rearrange the data, add a column of group 1(Biological Control) and 2(Chemical Control). Calculate the sample difference: $\overline{y}_1 - \overline{y}_2 = -24.875$.
Under $H_0 : \mu_1 = \mu_2$, randomly assign group 1 and 2.
Sample vector of size 16 of 2 types with the replacement, and repeat 10,000 times, count the number of times that the sample mean difference is as large as or larger than the observed sample mean difference -24.875, and compute a p-value as

$$2 \times \frac{8532}{10000} = 0.2974$$

Conclusion: There is no strong evidence against the H_0 that the mean stem volumn is the same for the two bud types. Accept the H_0 at the $\alpha = 0.05$ level.

(g) Since the data are paired two-sample, we employ Wilcoxon Signed Rank Test.
Firstly, we take the sample of differences and compute their absolute values. Then we sort the observations in the sample of absolute differences in ascending order and assign each observation a rank.

d_i	-4	-3	-6	-74	-10	-103	2	-1
$\|d_i\|$	4	3	6	74	10	103	2	1
rank	4	3	5	7	6	8	2	1

Add the ranks corresponding to the positive values in the original data set and this is the observed test statistic. In this case, rank of raw data is 2.
Perform a randomization test on this sum of the positive ranks. Repeat 10,000 times, count the number of times that the rank of random sample is as large as or smaller than the observed sample rank 2, and compute a p-value as

$$2 \times \frac{106}{10000} = 0.0212$$

Conclusion: There is strong evidence against the H_0 that the mean stem volumn is the same for the two bud types. Reject the H_0 at the $\alpha = 0.05$ level.

(h) From (b), the result from two sample T test is p-value = 0.1234, we then accept $H_0 : \mu_1 = \mu_2$. However, the normality assumption for this test are not all necessarily satisfied.
From (e), the result after transformation is p-value = 0.08882, we then still accept $H_0 : \mu_1 = \mu_2$.
From (f), the empirical p-value = 0.8532, we still accept $H_0 : \mu_1 = \mu_2$.
From (g), the empirical p-value = 0.0212, we then reject H_0 that the two populations have the same distribution, but accept H_1 that the two populations have the same shape but different locations.

Problem 6