Homework Assignment Number Six Solutions

We intend to purchase a liquid as a raw material for a material we are designing. Two vendors offer us samples of their product and a statistic sheet. We run the samples in our own labs and come up with the following data:

Vendor 1		Vendor 2	
sample #	outcome	sample #	outcome
1	2.30	1	2.49
2	2.49	2	1.98
3	2.05	3	2.18
4	2.40	4	2.36
5	2.18	5	2.47
6	2.12	6	2.36
7	2.38	7	1.82
8	2.39	8	1.88
9	2.40	9	1.87
10	2.46	10	1.87
11	4.19		
12	2.04		
13	2.43		
14	2.34		
15	2.19		
16	2.12		

Vendor Specification Claims:
Vendor 1: $\mu = 2.0$ and $\sigma^2 = 0.05$
Vendor 2: $\mu = 2.3$ and $\sigma^2 = 0.12$

Problem 1. (8 points)

Perform an outlier analysis on the data from both vendors.
(a) How many outliers are present in the data from vendor 1?
(b) Which data points from vendor 1 are outliers?
(c) How many outliers are present in the data from vendor 2?
(d) Which data points from vendor 2 are outliers?

Solution.

At the end of these solutions a script in Python is provided that was used to solve all parts of Homework 6 (and some additional confidence intervals not required in Homework 6). This script called six homemade functions, which are also reported in full at the end of the homework. After the source code, the code output is reported.

For Problem 1, we report extract the salient portion of the code output.

(a) How many outliers are present in the data from vendor 1?
(b) Which data points from vendor 1 are outliers?

```
   Step 2.1. identify outliers for data set #1
quartile one = 2.165
quartile three = 2.4074999999999998
interquartile range = 0.24249999999999972
lower outlier limit = 1.8012500000000005
upper outlier limit = 2.7712499999999993
Step 2.  Identify outliers
all data = 16
data 10 ([4.19]) is an outlier
good data = 15.0
The first data set has 15.0 data points.
data 10 ([4.19]) is an outlier
```

15 of the 16 data points of set #1 were good.
There was one outlier, the data point marked of 4.19.

(c) How many outliers are present in the data from vendor 2?
(d) Which data points from vendor 2 are outliers?

```
   Step 2.2. identify outliers for data set #2
quartile one = 1.8725
quartile three = 2.36
interquartile range = 0.4874999999999998
lower outlier limit = 1.1412500000000003
upper outlier limit = 3.0912499999999996
Step 2.  Identify outliers
all data = 10
good data = 10.0
The second data set has 10.0 data points.
```

All 10 data points of set #2 were good.
There were no outliers.

Problem 2. (8 points)

Compute the following statistics, excluding outliers.
(a) sample mean of vendor 1.
(b) sample standard deviation of vendor 1.
(c) sample mean of vendor 2.
(d) sample standard deviation of vendor 2.

Solution.

For Problem 2, we report extract the salient portion of the code output.
(a) sample mean of vendor 1.
(b) sample standard deviation of vendor 1.

```
   Step 3.1. determine statistics for data set #1
The mean is [2.286]
The standard deviation is [0.15426323]
```

(c) sample mean of vendor 2.
(d) sample standard deviation of vendor 2.

```
   Step 3.2. determine statistics for data set #2
The mean is [2.128]
The standard deviation is [0.27279622]
```

Problem 3. (4 points)

(a) Determine a 95% confidence interval on the mean of sample 1. Assume the given values of the population variances provided by the vendor are suspect and not to be trusted.
(b)Is the given population mean legitimate?

Solution.

(a) Determine a 95% confidence interval on the mean of sample 1. Assume the given values of the population variances provided by the vendor are suspect and not to be trusted.

For Problem 3, we report extract the salient portion of the code output.

```
P( [2.20057187] < mu < [2.37142813] ) is 0.95
```

$$v = n - 1 = 14$$

$$P(\bar{X} - t_\alpha \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_\alpha \frac{s}{\sqrt{n}}) = 1 - 2\alpha$$

$$P(2.20 < \mu < 2.37) = 0.95$$

(b)Is the given population mean legitimate?
No, the claimed population mean of 2.0 does not fall within a 95% confidence interval.

Problem 4. (4 points)

(a) Determine a 95% confidence interval on the difference of means between samples 1 and 2. Assume the given values of the population variances are suspect and not to be trusted.
(b) Is the difference between the given population means legitimate?

Solution.

(a) Determine a 95% confidence interval on the difference of means between samples 1 and 2. Assume the given values of the population variances are suspect and not to be trusted.

For Problem 4, we report extract the salient portion of the code output.

```
   Step 4.5. create confidence interval on difference of means
P( [-0.04748244] < dmu < [0.36348244] ) is 0.95
```

(b) Is the difference between the given population means legitimate?

$$v = \frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2/(n_1-1)\right]+\left[\left(\frac{s_2^2}{n_2}\right)^2/(n_2-1)\right]} \text{ if } \sigma_1 \neq \sigma_2$$

$$v = 12.87033673$$

confidence interval:

$$P\left[(\bar{X}_1 - \bar{X}_2) + t_\alpha\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < (\mu_1 - \mu_2) < (\bar{X}_1 - \bar{X}_2) + t_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right] = 1 - 2\alpha$$

$$P[-0.047 < (\mu_1 - \mu_2) < 0.363] = 0.95$$

(b) Is the difference between the given population means legitimate?

The claimed difference of population means of -0.30 does not fall within a 95% confidence interval.

Problem 5. (4 points)

(a) Determine a 95% confidence interval on the variance of sample 1.
(b) Is the given population variance legitimate?

Solution.

(a) Determine a 95% confidence interval on the variance of sample 1.
For Problem 5, we report extract the salient portion of the code output.

```
   Step 4.3. create confidence interval on variance for data set #1
P( [0.01275549] < sigma < [0.05918924] ) is 0.95
```

$$P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_\alpha}\right] = 1 - 2\alpha$$

$$P[0.0128 < \sigma^2 < 0.0592] = 0.95$$

(b) Is the given population variance legitimate?

Yes, the claimed population variance $\sigma^2 = 0.05$ falls within this interval.

Problem 6. (4 points)

(a) Determine a 95% confidence interval on the ratio of variance of samples 1 & 2.
(b) Is the ratio of the given population variances legitimate?

Solution.

(a) Determine a 95% confidence interval on the ratio of variance of samples 1 & 2.
For Problem 6, we report extract the salient portion of the code output.

```
   Step 4.6. create confidence interval on ratio of variances
P( [0.08419738] < ratiosig < [1.02626255] ) is 0.95
```

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha}(n_1-1,n_2-1)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha}(n_1-1,n_2-1)}\right] = 1-2\alpha$$

$$P\left[0.084 < \frac{\sigma_1^2}{\sigma_2^2} < 1.026\right] = 0.95$$

(b) Is the ratio of the given population variances legitimate?

Yes, the claimed ratio of population variances $\frac{\sigma_1^2}{\sigma_2^2} = \frac{0.05}{0.12} = 0.4167$ falls within this interval.

Python Codes

Main Script: hw6_v01.py

*** hw6_v01.py script start ***

```
# Homework 6
# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

import numpy as np
from find_outliers_funk import find_outliers_funk
from get_sample_stats_wo_outliers import get_sample_stats_wo_outliers
from confint_meanvarunknown_funk import confint_meanvarunknown_funk
from confint_var_funk import confint_var_funk
from confint_diffmeanvarunknown_funk import confint_diffmeanvarunknown_funk
from confint_rationvar_funk import confint_rationvar_funk

# Step 1.  Input Data
print("Step 1.  Input Data")

# Step 1.1. input data set #1
print("   Step 1.1. input data set #1")

datamat1 = np.array([
[2.30],
[2.49],
[2.05],
[2.40],
[2.18],
[2.12],
[2.38],
```

```
    [2.39],
    [2.40],
    [2.46],
    [4.19],
    [2.04],
    [2.43],
    [2.34],
    [2.19],
    [2.12],
    ])

    ndata1, nprop1 = datamat1.shape
    print(f"The first data set has a total of {ndata1} data points.")

    # Step 1.2. input dataset #2
    print("   Step 1.2. input data set #2")
    datamat2 = np.array([
    [2.49],
    [1.98],
    [2.18],
    [2.36],
    [2.47],
    [2.36],
    [1.82],
    [1.88],
    [1.87],
    [1.87],
    ])

    ndata2, nprop2 = datamat2.shape
    print(f"The first data set has a total of {ndata2} data points.")

    # Step 2. identify outliers
    print("Step 2. identify outliers")

    # Step 2.1. identify outliers for data set #1
    print("   Step 2.1. identify outliers for data set #1")
    ngood1, good1 = find_outliers_funk(datamat1)
    print(f"The first data set has {ngood1} data points.")
    for i in range(ndata1):
        if good1[i] == 0:
            print(f"data {i} ({datamat1[i]}) is an outlier")

    # Step 2.2. identify outliers for data set #2
    print("   Step 2.2. identify outliers for data set #2")
    ngood2, good2 = find_outliers_funk(datamat2)
    print(f"The second data set has {ngood2} data points.")
    for i in range(ndata2):
        if good2[i] == 0:
            print(f"data {i} ({datamat2[i]}) is an outlier")

    # Step 3.  determine statistics
    print("Step 3.  determine statistics")

    # Step 3.1. determine statistics for data set #1
    print("   Step 3.1. determine statistics for data set #1")
    mean_samp1, var_samp1, sd_samp1 = get_sample_stats_wo_outliers(datamat1, good1,
ngood1)

    # Step 3.2. determine statistics for data set #2
    print("   Step 3.2. determine statistics for data set #2")
    mean_samp2, var_samp2, sd_samp2 = get_sample_stats_wo_outliers(datamat2, good2,
ngood2)
```

```
    # Step 4.  create confidence intervals
    print("Step 4. create confidence intervals")

    # confidence interval
    confint = 0.95

    # Step 4.1. create confidence interval on mean for data set #1
    print("   Step 4.1. create confidence interval on mean for data set #1")
    ci_mulo1, ci_muhi1 = confint_meanvarunknown_funk(mean_samp1, sd_samp1, ngood1,
confint)

    # Step 4.2. create confidence interval on mean for data set #2
    print("   Step 4.2. create confidence interval on mean for data set #2")
    ci_mulo2, ci_muhi2 = confint_meanvarunknown_funk(mean_samp2, sd_samp2, ngood2,
confint)

    # Step 4.3. create confidence interval on variance for data set #1
    print("   Step 4.3. create confidence interval on variance for data set #1")
    ci_siglo1, ci_sighi1 = confint_var_funk(var_samp1, ngood1, confint)

    # Step 4.4. create confidence interval on variance for data set #2
    print("   Step 4.4. create confidence interval on variance for data set #2")
    ci_siglo2, ci_sighi2 = confint_var_funk(var_samp2, ngood2, confint)

    # Step 4.5. create confidence interval on difference of means
    print("   Step 4.5. create confidence interval on difference of means")
    dmu12lo, dmu12hi = confint_diffmeanvarunknown_funk(mean_samp1, var_samp1, ngood1,
mean_samp2, var_samp2, ngood2, confint)

    # Step 4.6. create confidence interval on ratio of variances
    print("   Step 4.6. create confidence interval on ratio of variances")
    ratiosig12lo, ratiosig12hi = confint_ratiovar_funk(var_samp1, ngood1, var_samp2,
ngood2, confint)
```

*** hw6_v01.py script end ***

This script calls six homemade functions. The source code for these functions is given below.

Find Outliers Function: find_outliers_funk.py

*** find_outliers_funk.py function start ***

```
import numpy as np

# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def find_outliers_funk(data):

    #print("Step 1.  Determine Outlier Limits")
    # find quartile one
    q1 = np.quantile(data, 0.25)
    print(f"quartile one = {q1}")
    # find quartile three
```

```
    q3 = np.quantile(data, 0.75)
    print(f"quartile three = {q3}")
    # interquartile range
    iqr = q3 - q1
    fac = 1.5
    print(f"interquartile range = {iqr}")
    #lower outlier limit
    ol_lo = q1 - fac*iqr
    print(f"lower outlier limit = {ol_lo}")
    #upper outlier limit
    ol_hi = q3 + fac*iqr
    print(f"upper outlier limit = {ol_hi}")
    # task 4:  what is the probability that data is good?

    print("Step 2.  Identify outliers")
    ndata = len(data)
    print(f"all data = {ndata}")
    # initialize a vector with 1's for good data
    good = np.ones(ndata)
    # insert 0's for outliers
    for i in range(len(data)):
        if data[i] < ol_lo or data[i] > ol_hi:
            good[i] = 0
            print(f"data {i} ({data[i]}) is an outlier")
    # number of good data
    ngood =sum(good)
    print(f"good data = {ngood}")

    return ngood, good
```

*** find_outliers_funk.py function end ***

Get Sample Statistics without Outliers Function: get_sample_stats_wo_outliers.py

*** get_sample_stats_wo_outliers.py function start ***

```
import numpy as np

# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def get_sample_stats_wo_outliers(datamat, good, ngood):

    ndata, nprop = datamat.shape
    mean_prop = 0.0
    mean_prop2 = 0.0

    # Compute mean and mean of squares explicitly
    for id in range(ndata):
        if good[id] == 1:
            mean_prop += datamat[id]
            mean_prop2 += datamat[id] ** 2
    mean_prop /= ngood
    mean_prop2 /= ngood
    print(f"The mean is {mean_prop}")
```

```
    # Compute variance and standard deviation explicitly
    var_prop = mean_prop2 - mean_prop ** 2
    var_prop = ngood/(ngood-1.0)*var_prop
    std_prop = np.sqrt(var_prop)
    print(f"The standard deviation is {std_prop}")

    return mean_prop, var_prop, std_prop
```

*** get_sample_stats_wo_outliers.py function end ***

Confidence Interval on Mean with Population Variance Unknown Function: confint_meanvarunknown_funk.py

*** get_confint_meanvarunknown_funk.py function start ***

```
import numpy as np
from scipy.stats import t

# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def confint_meanvarunknown_funk(sample_mean, sample_sd, n, confint):
    fac = sample_sd/np.sqrt(n)
    alpha = 0.5*(1.0 - confint)
    plo = alpha
    phi = 1.0 - alpha
    dof = n - 1
    tlo = t.ppf(plo, loc=0.0, df=dof)
#    print ({tlo})
    thi = t.ppf(phi, loc=0.0, df=dof)
    mulo = sample_mean + tlo*fac
    muhi = sample_mean + thi*fac
    print(f"P( {mulo} < mu < {muhi} ) is {confint}")
    return mulo, muhi
```

*** confint_meanvarunknown_funk.py function end ***

Confidence Interval on Variance Function: confint_var_funk.py

*** confint_var_funk.py function start ***

```
from scipy.stats import chi2
# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def confint_var_funk(sample_var, n, confint):
    alpha = 0.5*(1.0 - confint)
    plo = alpha
    phi = 1.0 - alpha
    dof = n - 1
    fac = dof*sample_var
    chi2lo = chi2.ppf(plo, df=dof)
    chi2hi = chi2.ppf(phi, df=dof)
```

```
        siglo = fac/chi2hi
        sighi = fac/chi2lo
        print(f"P( {siglo} < sigma < {sighi} ) is {confint}")

        return siglo, sighi
```

*** confint_var_funk.py function end ***

Confidence Interval on Difference of Means with Population Variance Unknown Function: confint_diffmeanvarunknown_funk.py

*** get_confint_diffmeanvarunknown_funk.py function start ***

```
import numpy as np
from scipy.stats import t

# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def confint_diffmeanvarunknown_funk(sample_mean1, sample_var1, n1,
sample_mean2, sample_var2, n2, confint):
        diffmeans = sample_mean1 - sample_mean2
        fac = np.sqrt( sample_var1/n1 + sample_var2/n2 )
        alpha = 0.5*(1.0 - confint)
        plo = alpha
        phi = 1.0 - alpha
        dof_numer = (sample_var1/n1 + sample_var2/n2)**2
        dof_denom1 = (sample_var1/n1)**2 / (n1 - 1)
        dof_denom2 = (sample_var2/n2)**2 / (n2 - 1)
        dof_denom = dof_denom1 + dof_denom2
        dof = dof_numer/dof_denom
        tlo = t.ppf(plo, df=dof)
        thi = t.ppf(phi, df=dof)
        dmulo = diffmeans + tlo*fac
        dmuhi = diffmeans + thi*fac
        print(f"P( {dmulo} < dmu < {dmuhi} ) is {confint}")
        return dmulo, dmuhi
```

*** confint_diffmeanvarunknown_funk.py function end ***

Confidence Interval on Ratio of Variances Function: confint_ratiovar_funk.py

*** confint_ratiovar_funk.py function start ***

```
from scipy.stats import f
# Author: David Keffer dkeffer@utk.edu
# Date begun: September 29, 2025
# Last updated: September 29, 2025

def confint_ratiovar_funk(sample_var1, n1, sample_var2, n2, confint):
    alpha = 0.5*(1.0 - confint)
    plo = alpha
    phi = 1.0 - alpha
    dof1 = n1 - 1
```

```
    dof2 = n2 - 1
#    print(f"{dof1} {dof2} {sample_var1} {sample_var2}")
    fac = sample_var1/sample_var2
    flo = f.ppf(plo, dfn=dof1, dfd=dof2)
    fhi = f.ppf(phi, dfn=dof1, dfd=dof2)
#    print({flo})
#    print({fhi})
    ratiosiglo = fac/fhi
    ratiosighi = fac/flo
    print(f"P( {ratiosiglo} < ratiosig < {ratiosighi} ) is {confint}")


    return ratiosiglo, ratiosighi
```

*** confint_ratiovar_funk.py function end ***

Code Output

This code is executed at the command line prompt:

```
run hw6_v01
```

It yields the following output:

```
Step 1.  Input Data
   Step 1.1. input data set #1
The first data set has a total of 16 data points.
   Step 1.2. input data set #2
The first data set has a total of 10 data points.
Step 2. identify outliers
   Step 2.1. identify outliers for data set #1
quartile one = 2.165
quartile three = 2.4074999999999998
interquartile range = 0.24249999999999972
lower outlier limit = 1.8012500000000005
upper outlier limit = 2.7712499999999993
Step 2.  Identify outliers
all data = 16
data 10 ([4.19]) is an outlier
good data = 15.0
The first data set has 15.0 data points.
data 10 ([4.19]) is an outlier
   Step 2.2. identify outliers for data set #2
quartile one = 1.8725
quartile three = 2.36
interquartile range = 0.4874999999999998
lower outlier limit = 1.1412500000000003
upper outlier limit = 3.0912499999999996
Step 2.  Identify outliers
all data = 10
good data = 10.0
The second data set has 10.0 data points.
Step 3.  determine statistics
   Step 3.1. determine statistics for data set #1
```

```
The mean is [2.286]
The standard deviation is [0.15426323]
   Step 3.2. determine statistics for data set #2
The mean is [2.128]
The standard deviation is [0.27279622]
Step 4. create confidence intervals
   Step 4.1. create confidence interval on mean for data set #1
P( [2.20057187] < mu < [2.37142813] ) is 0.95
   Step 4.2. create confidence interval on mean for data set #2
P( [1.93285334] < mu < [2.32314666] ) is 0.95
   Step 4.3. create confidence interval on variance for data set #1
P( [0.01275549] < sigma < [0.05918924] ) is 0.95
   Step 4.4. create confidence interval on variance for data set #2
P( [0.03520834] < sigma < [0.24802348] ) is 0.95
   Step 4.5. create confidence interval on difference of means
P( [-0.04748244] < dmu < [0.36348244] ) is 0.95
   Step 4.6. create confidence interval on ratio of variances
P( [0.08419738] < ratiosig < [1.02626255] ) is 0.95
```