Introduction to Essential Statistical Tests for Engineers with Python

Prepared by:

Dr. Gokhan Bingol (gbingol@hotmail.com)

October 18, 2024

Document version: 1.0

Updates will be available at: https://www.pebytes.com/pubs

Follow on YouTube: https://www.youtube.com/@sciencesuit

1. Introduction

There are a wide array of statistical tests each suited to specific data type, sample size and research question. Fundamental tools such as t-tests and ANOVA are commonly used for comparing means and analyzing variance, respectively (Montgomery, 2012). In addition to these, there are non-parametric tests like the Kruskal-Wallis or sign test, which are particularly useful when the data do not meet the normality assumptions (Gopal 2006; Kreyszig et al. 2011). Moreover, tests such as regression analysis allow further exploration of relationships between the variables (Montgomery *et al.* 2021).

Statistical tests and machine learning (ML) techniques are both valuable tools for process engineers and with the ever-rising popularity of ML process engineers need the necessary skills for descriptive & inferential statistics and supervised/unsupervised modeling methods (Pinheiro & Patetta, 2021). Statistical tests are well-established methods that provide interpretable results and clear hypotheses testing (Montgomery 2012) and are particularly useful for determining the significance of relationships when working with small datasets (Box *et al*. 2005). On the other hand, ML techniques excel in handling complex datasets with intricate relationships that may not be easily captured by traditional statistics (Hastie *et al*. 2009); however, they require more data and computational resources and their "black-box" nature, particularly deep learning methods, might lack the interpretability needed in process engineering decisions (Rudin 2019)[1].

Given the vast variety of statistical methods available, the current edition focuses on the fundamental statistical tests most relevant for process engineers. The target audience of the current work is engineers and therefore this document assumes the reader already has some background in calculus, statistics and statistical distributions. Furthermore, at least a basic level of understanding of Python is required.

This work heavily uses the following Python packages: *numpy* and *scisuit*[2]. The design of *scisuit*'s statistical library is inspired by *R*[3] and therefore the knowledge gained from the current work can be conveniently adapted to *R*, which is a popular software in data science realm.

---

1 **Rudin C** (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206-215.

2 **scisuit** at least v1.4.0. Unless otherwise stated, alternate name **np** is used for **numpy**

3 https://www.r-project.org

2. Fundamentals

2.1. Point Estimation

A point estimate is a single value (i.e., mean, median, proportion, ...) based on sampled data to represent a plausible value of a population characteristics (Peck *et al*. 2016).

Example 2.1

Out of 7421 US College students 2998 reported using internet more than 3 hours a day. What is the proportion of *all* US College students who use internet more than 3 hours a day? (Adapted from Peck et al. 2016).

Solution:

The solution is straightforward: $p = \frac{2998}{7421} = 0.40$

Based on the statistics it is possible to claim that approximately 40% of the students in US spend more than 3 hours a day using the internet. Please note that based on the survey result, we made a claim about the population, students in US. ■

Now that we made an estimate based on the survey, we should ask ourselves: "*How reliable is this estimate?*". We know that if we had another group of students, the percentage might not have been 40, maybe it would be 45 or 35. There are no perfect estimators but we expect that *on average* the estimator should gives us the right answer.

2.1.1. Unbiased estimators

Definition: A statistic Θ is an unbiased estimator of the parameter θ of a given distribution if and only if,

$$E(\Theta) = \theta \qquad (2.1)$$

for all possible values of θ (Miller and Miller 2014).

Example 2.2

If X has binomial distribution with the parameters *n* and *p*, show that the sample proportion, *X/n* is an unbiased estimator of *p*.

Before we proceed with the solution, let's refresh ourselves with a simple example: Suppose we conduct an experiment where we flip a coin 10 times. We already know that the probability of getting heads (success) is *p*=0.5. However, we want to estimate *p* by flipping the coin and calculating the sample proportion, *X/n*. If we flip the coin 10 times and get *X=6* heads, *p=0.6*. However, after many experiments *p* will be found as 0.5. Therefore, *X/n* is an unbiased estimator.

$$E\left(\frac{X}{n}\right)=\frac{1}{n}E(X)=\frac{1}{n}\cdot np=p$$

therefore, *X/n* is an unbiased estimator of *p*. ■

Example 2.3

Prove that $E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$ is unbiased estimator of population variance (σ^2).

Solution:

$$E(S^2)=E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]$$

Now we are going to add and subtract μ inside the parenthesis:

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i-\mu)-(\bar{X}-\mu)\right)^2\right]$$

After a straightforward algebraic manipulation,

$$=\frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i-\mu)^2-(\bar{X}-\mu)^2\right)\right]$$

$$=\frac{1}{n-1}\left[\sum_{i=1}^{n}E(X_i-\mu)^2-n(\bar{X}-\mu)^2\right]$$

Note that $E(X_i-\mu)^2=\sigma^2$ and $E(\bar{X}-\mu)^2=\frac{\sigma^2}{n}$. Putting the knowns in the last equation,

$$=\frac{1}{n-1}\left(n\cdot\sigma^2-n\cdot\frac{\sigma^2}{n}\right)=\sigma^2$$

Therefore, $E(S^2)$ is an unbiased estimator of population variance. ■

There are other properties of estimators: **i)** minimum variance, **ii)** efficiency, **iii)** consistency, **iv)** sufficiency, and **v)** robustness. Interested readers can refer to textbooks on mathematical statistics (Devore *et al*. 2021; Larsen & Marx 2011; Miller & Miller 2014).

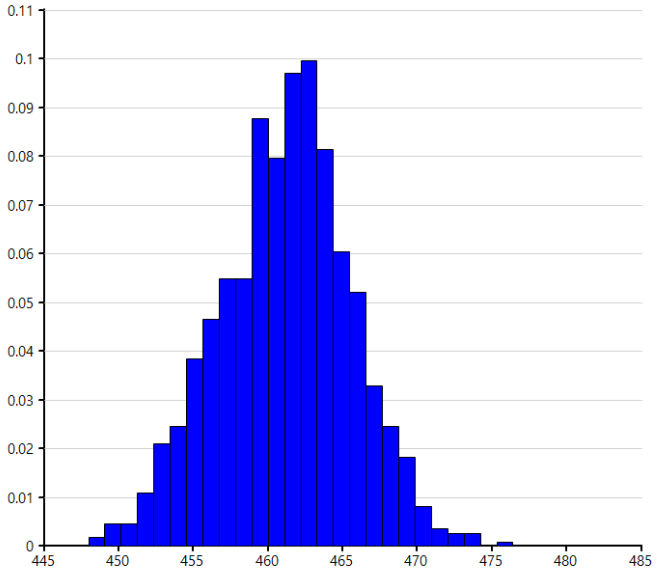
2.2. Statistical Confidence

Suppose you want to estimate the SAT scores of students. For that purpose, a randomly selected 500 students have been given an SAT test and a mean value of 461 is obtained (adapted from Moore *et al*. 2009). Although it is known that the sample mean is an *unbiased* estimator of the population mean (μ), we already know that had we sampled another 500 students, the mean could (most likely would) have been different than 461. Therefore, how confident are we to claim that the population mean will be 461.

Suppose that the standard deviation of the population is known (σ=100). We know that if we repeat sampling 500 samples, the mean of these samples will follow the N(μ, $\frac{100}{\sqrt{500}}=4.5$) curve. Let's demonstrate this will a simple script:

Script 2.1

```
import scisuit.plot as plt
from scisuit.stats import rnorm
aver = []
for i in range(1000):
    sample = rnorm(n=500, mean= 461, sd= 100)
    aver.append(sum(sample)/500)

plt.hist(data=aver, density=True)
plt.show()
```



It is seen that the interval (447.5, 474.5) represents almost all possible mean values. Therefore we are 99.7% (3σ) confident (confidence level) that the population mean will be in this interval. Note also that, as a natural consequence our confidence level decreases as the interval length decreases.

$$461-3\times4.5=447.5$$

$$461+3\times4.5=474.5$$

Fig 2.1: *Density scaled histogram*

2.3. Confidence Intervals

A way to quantify the amount of uncertainty in a point estimator is to construct a confidence interval (Larsen & Marx 2011). The definition of confidence interval is as follows: "*... an interval computed from sample data by a method that has probability C of producing an interval containing the true value of the parameter.*" (Moore *et al*. 2009). Peck *et al*. (2016) gives a general form of confidence interval as follows:

$$\begin{pmatrix}\text{point estimate using a}\\ \text{specified statistic}\end{pmatrix}\pm(\text{critical value})\cdot\begin{pmatrix}\text{estimated standard deviation}\\ \text{of the statistic}\end{pmatrix} \qquad \textbf{(2.2)}$$

Note that the estimated standard deviation of the statistic is also known as *standard error*. In other words, when the standard deviation of a statistic is estimated from the data (because the population's standard deviation is not known), the result is called the standard error of the statistic (Moore *et al*. 2009).

Example 2.4

Establish a confidence interval for binomial distribution.

Solution:

We already know that Abraham DeMoivre showed that when X is a binomial random variable and n is large the probability can be approximated as follows:

$$\lim_{n\to\infty} P\left(a \le \frac{X-np}{\sqrt{np(1-p)}} \le b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-z^2/2} dz$$

To establish an approximate 100(1−α)% confidence interval,

$$P\left[-z_{\alpha/2} \le \frac{X-np}{\sqrt{np(1-p)}} \le z_{\alpha/2}\right] = 1-\alpha$$

Rewriting the equation,

$$P\left[-z_{\alpha/2} \le \frac{X/n-p}{\sqrt{\frac{(X/n)(1-X/n)}{n}}} \le z_{\alpha/2}\right] = 1-\alpha$$

Rewriting the equation by isolating p leads to,

$$\left[\frac{k}{n} - z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}}, \frac{k}{n} + z_{\alpha/2}\sqrt{\frac{(k/n)(1-k/n)}{n}},\right]$$

■

If a 95% confidence interval to be established, then $z_{\alpha/2}$ would be ≈1.96.

Script 2.2

```
alpha1 = 0.05
alpha2 = 0.01
print(qnorm(alpha1/2), qnorm(1-alpha1/2))
print(qnorm(alpha2/2), qnorm(1-alpha2/2))
```

```
-1.95996        1.95996
-2.57583        2.57583
```

Note that if a 95% confidence interval (CI) yields an interval (0.52, 0.57), it is *tempting* to say that there is a probability of 0.95 that p will be in between 0.52 and 0.57. Larsen & Marx (2011) and Peck *et al*. (2016) warns against this *temptation*. A close look at Eq. (2.2) reveals that from sample to sample

the constructed CI will be different. However, *in the long* run 95% of the constructed CIs will contain the true *p* and 5% will not. This is well depicted in the figure (Figure 9.4 at pp. 471) presented by Peck *et al*. (2016).

Note also that a 99% CI will be wider than a 95% CI. However, the higher reliability causes a loss in precision. Therefore, Peck *et al*. (2016) remarks that many investigators consider a 95% CI as a reasonable compromise between reliability and precision.

2.4. Hypothesis Testing

Confidence intervals and statistical tests are the two most important ideas in the age of modern statistics (Kreyszig *et al*. 2011). The confidence interval is carried out when we would like to estimate population parameter. Another type of inference is to assess the evidence provided by data against a claim about a parameter of the population (Moore *et a*l. 2009). Therefore, after carrying out an experiment conclusions must be drawn based on the obtained data. The two competing propositions are called the *null hypothesis* (H_0) and the *alternative hypothesis* (H_1) (Larsen & Marx 2011).

We initially assume that a particular claim about a population (H_0) is correct. Then based on the evidence from data we either reject **H_0** and accept **H_1** if there is *compelling* evidence or accept **H_0** in favor of **H_1** (Peck *et al*. 2016).

An example from Larsen & Marx (2011) would clarify the concepts better: Imagine as an automobile company you are looking for additives to increase gas mileage. Without the additives, the cars are known to average 25.0 mpg with a σ=2.4 mpg and with the addition of additives, it was found (experiment involved 30 cars) that the mileage increased to 26.3 mpg.

Now, in terms of null and alternative hypothesis, **H_0** is 25 mpg and **H_1** claims 26.3 mpg. We know that if the experiments were carried out with another 30 cars, the result would be different (lower or higher) than 26.3 mpg. Therefore, "*is an increase to 26.3 mpg due to additives or not?*". At this point we should rephrase our question: "*if we sample 30 cars from a population with μ=25.0 mpg and σ=2.4, what are the chances that we will get 26.3 mpg on average?*". If the chances are high, then the additive is not working; however, if the chances are low, then it must be due to the additives that the cars are getting 26.3 mpg.

Let’s evaluate this with a script (note the similarity to Script 2.1):

Script 2.3

```
aver = []
for i in range(10000):
   sample = rnorm(n=30, mean= 25, sd= 2.4)
   aver.append(sum(sample)/30)

filtered = list(filter(lambda x: x>=26.5, aver))
print(f"probability = {len(filtered)/len(aver)}")
```

```
probability = 0.0002
```

We observe that the probability is too low for this to happen by chance (random sampling from the population). Therefore, we conclude that in light of the statistical evidence the additives indeed work (H_1 wins) and reject H_0.

Directly computing the probability:

$$P\left(\frac{26.50-25.0}{2.4/\sqrt{30}}\right)=0.0003$$

which is very close to the simulation result of Script (2.3).

Wackerly *et al*. (2008) lists the elements of a statistical test as follows:

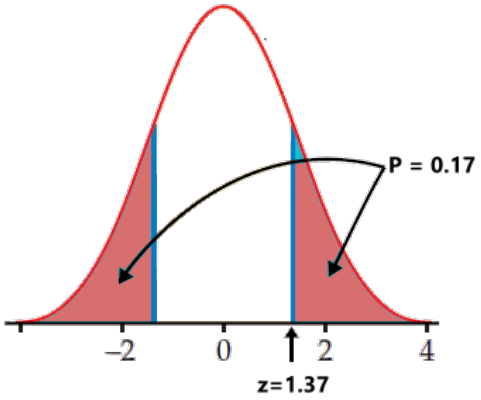
1. Null hypothesis (μ=25.0 mpg),
2. Alternative hypothesis (26.5 mpg),
3. Test statistic ($\frac{\bar{x}-25.0}{2.4/\sqrt{30}}$),
4. Rejection region ($\bar{x}\geq 25.718$ for $\alpha=0.05$)

■

2.4.1. The P-Value

We have seen that using a level of significance a critical region (H_0 being rejected) can be identified (Larsen & Marx 2011); for example, $z \geq 2$ is a rejection criteria for the Supreme Court of the United States (Moore et al. 2009). However, not all test statistics are normal and therefore a new strategy is to calculate the p-value, which is defined as (Larsen & Marx 2011): "*… the probability of getting a value for that test statistic as extreme as or more extreme than what was actually observed given that H_0 is true.*". The term *extreme* is also used by Moore *et al*. (2009) and is explained as "*far from what we would expect if H_0 were true*".

If for example the test statistics yield $Z=1.37$ and we are carrying out a two-sided test, the p-value would be, $P(Z \leq -1.37 \text{ or } Z \geq 1.37)$ where Z has a standard normal distribution.




```
z=1.37

#pnorm computes left-tailed probability
pvalue = pnorm(-z) + (1-pnorm(z))

print(f"p-value = {pvalue}")
```

```
p-value = 0.17
```

Fig 2.2: *P-value (i.e. $z \geq 1.37$ is considered extreme) (adapted from Moore et al. 2009)*

A simpler definition is given by Miller & Miller (2014): "*… the lowest level of significance at which the null hypothesis could have been rejected*". Let's rephrase Miller & Miller (2014) definition: once a level of significance is decided (e.g. $\alpha=0.05$), if the computed *p-value* is less than the α, then we reject H_0. For example, in the gasoline additive example, p-value was computed as 0.0003 and if $\alpha=0.05$, then since $p < \alpha$, we reject H_0 in favor of H_1 (i.e., additive has effect).

In terms of a standard normal distribution, there are 3 cases of computing p-values:

1. $H_1: \mu > \mu_0 \rightarrow P(Z \geq z)$ (alternative is greater than)
2. $H_1: \mu < \mu_0 \rightarrow P(Z \leq z)$ (alternative is smaller than)
3. $H_1: \mu \neq \mu_0 \rightarrow 2P(Z \geq |z|)$ (alternative is not equal)

Example 2.5

A bakery claims on its packages that its cookies are 8 g. It is known that the standard deviation of the 8 g packages of cookies is 0.16 g. As a quality control engineer, you collected 25 packages and found that the average is 8.091 g. Is the production process going alright? (adapted from Miller & Miller 2014).

Solution:

The null hypothesis is H_0: μ=8 g,

The alternative hypothesis H_1: μ≠8 g.

The test statistic: $z = \frac{8.091 - 8}{0.16/\sqrt{25}} = 2.84$

```
>> 1-pnorm(2.84) + pnorm(-2.84) #2*(1- pnorm(2.84))
0.0045
```

Since $p<0.05$, we reject the null hypothesis. Therefore, the process should be checked and suitable adjustments should be made. ■

3. Z-Test for Population Means

The fundamental limitation to applying z-test is that the population variance must be known in advance (Kanji 2006; Moore *et al*. 2009; Peck *et al*. 2016). The test is accurate when the population is normally distributed; however, it will give an approximate value even if the population is not normally distributed (Kanji 2006). In most practical applications, population variance is unknown and the sample size is small therefore a *t-test* is more commonly used.

3.1. One-sample z-test

From a population with known mean (μ) and variance (σ), a random sample of size *n* is taken (generally n≥30) and the sample mean ($\bar{x}$) calculated. The test statistic:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} \tag{3.1}$$

Example 3.1

A filling process is set to fill tubs with powder of 4 g on average. For this filling process it is known that the standard deviation is 1 g. An inspector takes a random sample of 9 tubs and obtains the following data:

Weights = [3.8, 5.4, 4.4, 5.9, 4.5, 4.8, 4.3, 3.8, 4.5]

Is the filling process working fine? (Adapted from Kanji 2006).

Solution:

The average of 9 samples is: 4.6 g

Test statistic: $Z = \frac{4.6 - 4}{1/\sqrt{9}} = 1.8,$

Since 1.8 is in the range of -1.96<Z<1.96, we cannot reject the null hypothesis, therefore the filling process works fine (i.e. there is no evidence to suggest it is different than 4 g).

Is it over-filling?

Now, we are going to carry out 1-tailed z-test and therefore acceptance region is Z<1.645. Since the test statistic is greater than 1.645, we reject the null hypothesis and have evidence that the filling process is over-filling.

Script 3.1

```
import scisuit.plot as plt
from scisuit.stats import test_z

data = [3.8, 5.4, 4.4, 5.9, 4.5, 4.8, 4.3, 3.8, 4.5]
result = test_z(x=data, sd1=1, mu=4)
print(result)
```

```
N=9, mean=4.6, Z=1.799
p-value = 0.072 (two.sided)
Confidence interval (3.95, 5.25)
```

Since $p>0.05$, we cannot reject H_0.

Script 3.1 requires minor change to analyze whether it is over-filling or not. We will set the parameter, namely *alternative*, to “greater” whose default value was set to “two.sided”.

Script 3.2

```
result = test_z(x=data, sd1=1, mu=4, alternative="greater")
print(result)
```

```
p-value = 0.036 (greater)
Confidence interval (4.052, inf)
```

Since $p<0.05$, we reject the null hypothesis in favor of alternative hypothesis.

3.2. Two-sample z-test

In essence, two-sample is very similar to one-sample z-test such that we take n_1 and n_2 samples from two populations with means (μ_1 and μ_2) and variances (σ_1 and σ_2). Therefore, the test statistic is computed as:

$$Z=\frac{\left(\bar{x}_1-\bar{x}_2\right)-\left(\mu_1-\mu_2\right)}{\left(\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}\right)^{\frac{1}{2}}} \qquad \textbf{(3.2)}$$

Example 3.2

A survey has been conducted to see if studying over or under 10 h/week has an effect on overall GPA. For those who studied less (x) and more (y) than 10 h/week the GPAs were:

x=[2.80, 3.40, 4.00, 3.60, 2.00, 3.00, 3.47, 2.80, 2.60, 2.0]

y = [3.00, 3.00, 2.20, 2.40, 4.00, 2.96, 3.41, 3.27, 3.80, 3.10, 2.50].

respectively. It is known that the standard deviation of GPAs for the whole campus is σ=0.6. Does studying over or under 10 h/week has an effect on GPA? (Adapted from Devore et al. 2021)

Solution:

We have two groups (those studying over and under 10 h/week) from the same population (whole campus) whose standard deviation is known (σ=0.6).

We will solve this question directly using a Python script and the mathematical computations are left as an exercise to the reader.

Script 3.3

```
x = [2.80, 3.40, 4.00, 3.60, 2.00, 3.00, 3.47, 2.80, 2.60, 2.0]
y = [3.00, 3.00, 2.20, 2.40, 4.00, 2.96, 3.41, 3.27, 3.80, 3.10, 2.50]
mu = 0
sd1, sd2 = 0.6, 0.6

result = test_z(x=x, y=y, sd1=sd1, sd2=sd2, mu=0)
print(result)
```

```
n1=10, n2=11, mean1=2.967, mean2=3.058
Z=-0.3478
p-value = 0.728 (two.sided)
Confidence interval (-0.605, 0.423)
```

Since $p>0.05$ there is no statistical evidence to reject H_0 ($\mu_1 - \mu_2 = 0$) and therefore there is no statistically significant different between studying over or under 10 h/week.

4. Student's t-test for Population Means

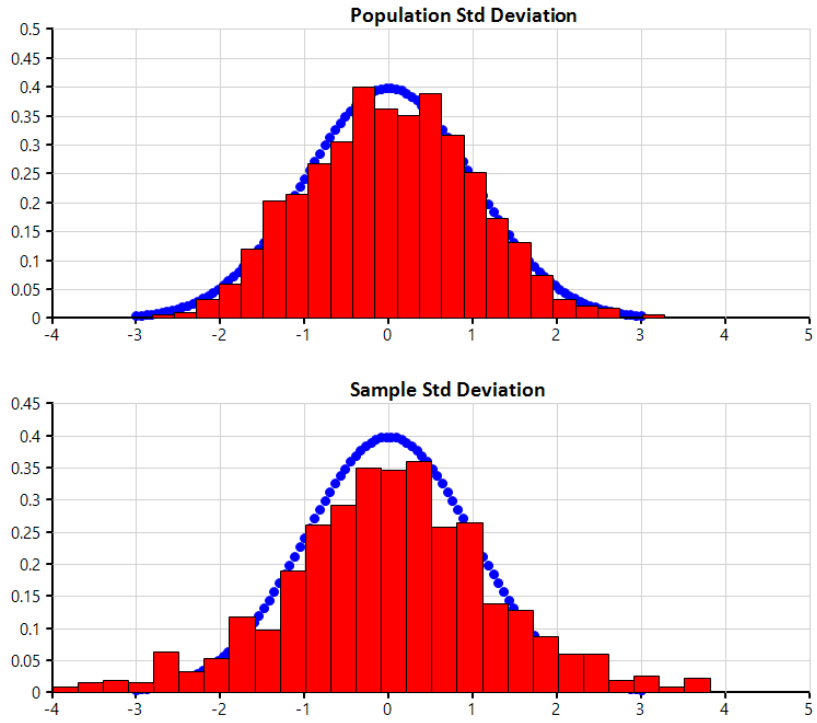
In Chapter 3, we have seen that *z-test* is only possible when standard deviation (σ) of the population is known. Therefore, the z-statistic is not commonly used (Peck *et al*. 2016). When standard deviation(s) are not known they must be estimated from samples and in Example (2.3) it was already shown that S^2 is an unbiased estimator of σ^2. Therefore, one might immediately be *tempted* to use Eq. (3.1).

Then, comes the question: *What effect does replacing σ with S have on Z ratio*? (Larsen & Marx 2011). In order to answer this question, let's demonstrate the effect of replacing σ with S on Z ratio with a script:

Script 4.1

```
import numpy as np
import scisuit.plot as plt
from scisuit.stats import dnorm, rnorm

#plotting f(z) curve
x = np.linspace(-3, 3, num=100)
y = dnorm(x)

N = 4
sigma, mu = 1.0,  0.0 #stdev and mean of population
z, t = [], []
for i in range(1000):
   sample = rnorm(n=N)
   aver = sum(sample)/N

   #using population stdev
   z_ratio = (aver-mu)/(sigma/sqrt(N))
   z.append(z_ratio)

   #computing stdev from sample
   s = float(np.std(sample, ddof=1))
   z_ratio = (aver-mu)/(s/sqrt(N))

   #filter out too big and too small ones
   if(-4<z_ratio<4):
      t.append(z_ratio)

plt.layout(nrows=2, ncols=1)

plt.subplot(row=0, col=0)
plt.scatter(x=x, y=y)
```

```
plt.hist(data=z, density=True)
plt.title("Population Std Deviation")

plt.subplot(row=1, col=0)
plt.scatter(x=x, y=y)
plt.hist(data=t, density=True)
plt.title("Sample Std Deviation")

plt.show()
```



In the top figure, it is seen that when the standard deviation of the population (σ) is known *f(z)* is consistent with $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.

However, when σ is not known and instead *S* is used to compute z-ratio, $\frac{\bar{x}-\mu}{S/\sqrt{n}}$, it is seen that *f(z)* underestimates the ratios much less than zero as well as the ratios much larger than zero.

Credit for recognizing this difference goes to *William Sealy Gossett*.[4]

Fig 4.1: $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ *(top) vs* $\frac{\bar{x}-\mu}{S/\sqrt{n}}$

Note that in Script (4.1), *N* was intentionally chosen a small value (*N=4*). It is recommended to change *N* to a greater number, such as 10, 20 or 50 in order to observe the effect of large samples.

4 **Student** (1908). The probable error of a mean. Biometrika, 6(1), 1-25.

4.1. One-sample t-test

Let $\bar{x}$ and s be the mean and standard deviation of a random sample from a normally distributed population. Then,

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}} \tag{4.1}$$

has a *t* distribution with *df=n-1*. Here s is the sample's standard deviation and computed as:

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \tag{4.2}$$

Example 4.1

In 2006, a report revealed that UK subscribers with 3G phones listen on average 8.3 hours/month full-track music. The data for a random sample of size 8 for US subscribers is x=[5, 6, 0, 4, 11, 9, 2, 3]. Is there a difference between US and UK subscribers? (Adapted from Moore et al. 2009).

Solution:

Script 4.2

```
from statistics import stdev
from scisuit.stats import qt

x=[5, 6, 0, 4, 11, 9, 2, 3]
n = len(x)
df = n-1 #degrees of freedom
aver = sum(x)/n
stderr = stdev(x)/sqrt(n) #standard error

#construct a 95% interval
tval = qt(0.025, df=df) #alpha/2=0.025
v_1 = aver - tval*stderr
v_2 = aver + tval*stderr
print(f"Interval: ({min(v_1, v_2)}, {max(v_1, v_2)})")
```

```
Interval: (1.97, 8.03)
```

Since the confidence interval does not contain 8.3 and furthermore since its upper limit is smaller than 8.3, it can be concluded that US subscribers listen less than UK subscribers.

Directly solving using *scisuit*'s built-in function:

Script 4.3

```
from scisuit.stats import test_t
x=[5, 6, 0, 4, 11, 9, 2, 3]
result = test_t(x=x, mu=8.3)
print(result)
```

```
One-sample t-test for two.sided
N=8, mean=5.0
SE=1.282, t=-2.575
p-value =0.037
Confidence interval: (1.97, 8.03)
```

Since $p<0.05$ we reject H_0 and claim that there is statistically significant difference between US and UK subscribers. [If in `test_t` function H_1 was set to "less" instead of "two.sided" then $p=0.018$. Therefore, we would reject the H_0 in favor of H_1, i.e. US subscribers indeed listen less than UK's.]

4.2. Two-sample t-test

4.2.1. Equal Variances

Assume we are drawing n and m samples from two populations, namely X and Y, with equal variances, s^2, but with different means μ_1 *and* μ_2. Let S_P^2 be the pooled variance, then:

$$S_P^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n+m-2} \tag{4.3}$$

and test statistic is defined as:

$$T_{n+m-2}=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{S_p\sqrt{\frac{1}{n}+\frac{1}{m}}} \tag{4.4}$$

has a Student's t-distribution with $n+m-2$ degrees of freedom.

Example 4.2

Student surveys are important in academia. An academic who scored low on a student survey joined workshops to improve "enthusiasm" in teaching. X and Y are survey scores from his fall and spring semester classes which he selected to *have the same demographics*.

X= [3, 1, 2, 1, 3, 2, 4, 2, 1]

Y = [5, 4, 3, 4, 5, 4, 4, 5, 4]

Is there a difference in scores of both semester? (Adapted from Larsen & Marx 2011).

Solution:

We can make the following assumptions:

1. The variance of the populations are not known, therefore z-test cannot be applied.
2. It is reasonable to assume equal variances since the X and Y have the same demographics.

Script 4.4

```
from scisuit.stats import test_t
x = [3, 1, 2, 1, 3, 2, 4, 2, 1]
y = [5, 4, 3, 4, 5, 4, 4, 5, 4]
result = test_t(x=x, y=y, varequal=True)
print(result)
```

```
Two-sample t-test assuming equal variances
n1=9, n2=9, df=16
s1=1.054, s2=0.667
Pooled std = 0.882
t = -5.07
p-value = 0.0001 (two.sided)
Confidence interval: (-2.992, -1.230)
```

Since $p<0.05$, the difference between the scores of fall and spring are statistically significant.

4.2.2. Unequal Variances

Similar to section 4.2.1, we are drawing random samples of size n_1 and n_2 from normal distributions with means μ_X and μ_Y, but with standard deviations σ_X and σ_Y, respectively.

$$S_1^2=\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n_1-1} \text{ and } S_2^2=\frac{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}{n_2-1} \quad \textbf{(4.5)}$$

The test statistic is computed as follows:

$$t=\frac{\bar{X}-\bar{Y}-(\mu_X-\mu_Y)}{\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}} \quad \textbf{(4.6)}$$

In 1938 Welch[5] showed that t is approximately distributed as a Student's t random variable with df:

$$df=\frac{\left(\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)}+\frac{s_2^4}{n_2^2(n_2-1)}} \quad \textbf{(4.7)}$$

Example 4.3

A study by Larson and Morris[6] (2008) surveyed the annual salary of men and women working as purchasing managers subscribed to *Purchasing* magazine. The salaries are (in thousands of US dollars):

Men = [81, 69, 81, 76, 76, 74, 69, 76, 79, 65]

Women = [78, 60, 67, 61, 62, 73, 71, 58, 68, 48]

Is there a difference in salaries between men and women? (Adapted from Peck et al. 2016)

5 https://www.jstor.org/stable/2332010

6 **Larson PD & Morris M** (2008). Sex and Salary: A Survey of Purchasing and Supply Professionals, Journal of Purchasing and Supply Management, 112–124.

Solution:

Following assumption can be made:

1. Z-test cannot be applied because the variance of the populations are not known.
2. Although the samples were selected from the subscribers of *Purchasing* magazine, Larson and Morris (2008) considered two populations of interest, i.e. male and female purchasing managers. Therefore, equal variances should not be applied.

Script 4.5

```
from scisuit.stats import test_t
Men = [81, 69, 81, 76, 76, 74, 69, 76, 79, 65]
Women = [78, 60, 67, 61, 62, 73, 71, 58, 68, 48]
result = test_t(x=Women, y=Men, varequal=False)
print(result)
```

```
Two-sample t-test assuming unequal variances
n1=10, n2=10, df=15
s1=8.617, s2=5.399
t = -3.11
p-value = 0.007 (two.sided)
Confidence interval: (-16.7, -3.1)
```

Since p<0.05, there is statistically significant difference between salaries of each group.

4.3. Paired t-test

In essence a paired t-test is a two-sample t-test as there are two samples. However, the two samples are *not independent* as one of the factors in the first sample is paired in a meaningful way with a particular observation in the second sample (Larsen & Marx 2011; Peck et al. 2016).

The equation to compute the test statistics is similar to one-sample t-test, Eq. (4.1):

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \tag{4.8}$$

where $\bar{x}$ and s are mean and standard deviation of the sample differences, respectively. The degrees of freedom is: *df=n-1*.

Example 4.4

In a study where 6th grade students who had not previously played chess participated in a program in which they took chess lessons and played chess daily for 9 months. Below data demonstrates their memory test score before and after taking the lessons:

Pre = [510, 610, 640, 675, 600, 550, 610, 625, 450, 720, 575, 675]

Post = [850, 790, 850, 775, 700, 775, 700, 850, 690, 775, 540, 680]

Is there evidence that playing chess increases the memory scores? (Adapted from Peck *et al*. 2016).

Solution:

Before we attempt to solve the question, we make the following assumptions:

1. Z-test cannot be applied since population variance is not known,
2. Pre- and post-test scores are not independent since they were applied to the same subjects.

Script 4.6

```
from scisuit.stats import test_t
Pre = [510, 610, 640, 675, 600, 550, 610, 625, 450, 720, 575, 675]
Post = [850, 790, 850, 775, 700, 775, 700, 850, 690, 775, 540, 680]
result = test_t(x=Post, y=Pre, paired=True)
print(result)
```

```
Paired t-test for two.sided
N=12, mean1=747.9, mean2=603.3, mean diff=144.6
t =4.564
p-value =0.0008
Confidence interval: (74.9, 214.3)
```

Since $p<0.05$, there is statistical evidence that playing chess indeed made a difference in increasing the memory scores.

If the parameter, namely *alternative*, was set to "less", then $p=0.99$. Therefore, we would reject the alternative hypothesis (*Post<Pre*). However, on the other hand, alternative was set to "greater" then $p=0.0004$, therefore we would reject the H_0 and accept H_1 (*Post>Pre*).

5. F-Test for Population Variances

Assume that a metal rod production facility uses two machines on the production line. Each machine produces rods with thicknesses μ_X and μ_Y which are not significantly different. However, if the variabilities are significantly different, then some of the produced rods might become unacceptable as they will be outside the engineering specifications.

In Section (4.2), it was shown that there are two cases for two-sample *t-tests*: whether variances were equal or not. To be able to choose the right procedure, Larsen & Marx (2011) recommended that *F test* should be used prior to testing for $\mu_X=\mu_Y$.

Let's draw random samples from populations with normal distribution. Let $X_1, \dots, X_m$ be a random sample from a population with standard deviation σ_1 and let $Y_1, \dots, Y_n$ be another random sample from a population with standard deviation σ_2. Let S_1 and S_2 be the sample standard deviations. Then the test statistic is:

$$F=\frac{S_1^2/\sigma_1}{S_2^2/\sigma_2} \qquad \textbf{(5.1)}$$

has an F distribution with $df_1=m-1$ and $df_2=n-1$, (Devore et al. 2021).

Example 5.1

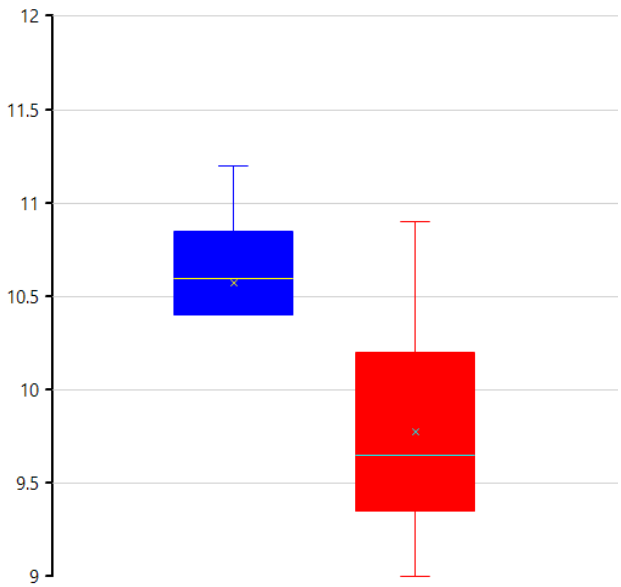
α-waves produced by brain have a characteristic frequency from 8 to 13 Hz. The subjects were 20 inmates in a Canadian prison who were randomly split into two groups: one group was placed in solitary confinement; the other group was allowed to remain in their own cells. Seven days later, α-wave frequencies were measured for all twenty subjects are shown below:

non-confined = [10.7, 10.7, 10.4, 10.9, 10.5, 10.3, 9.6, 11.1, 11.2, 10.4]
confined = [9.6, 10.4, 9.7, 10.3, 9.2, 9.3, 9.9, 9.5, 9, 10.9]

Is there a significant difference in variability between two groups?

Solution:

Using a box-whisker plot, let's first visualize the data as shown in Fig. (5.1).



It is seen that inmates placed in solitary confinement (red box) show a clear decrease in the α-wave frequency.

Furthermore, the variability of that particular group seems higher than non-confined inmates.

Fig 5.1: *Non-confined (blue) vs solitary confined (red)*

Script 5.1

```
from scisuit.stats import test_f, test_f_Result

nonconfined = [10.7, 10.7, 10.4, 10.9, 10.5, 10.3, 9.6, 11.1, 11.2, 10.4]
confined = [9.6, 10.4, 9.7, 10.3, 9.2, 9.3, 9.9, 9.5, 9, 10.9]
result = test_f(x=confined, y=nonconfined)
print(result)
```

```
F test for two.sided
df1=9, df2=9, var1=0.357, var2=0.211
F=1.696
p-value =0.443
Confidence interval: (0.42, 6.83)
```

Since $p>0.05$, we cannot reject H_0 ($\sigma_1=\sigma_2$). Therefore, there is no statistically significant difference between the variances of two groups.

6. Analysis of Variance (ANOVA)

In Section (4.2) we have seen that when exactly two means needs to be compared, we could use two-sample t-test. The methodology for *comparing several means* is called analysis of variance (ANOVA). When there is only a single factor with multiple levels, i.e. color of strawberries subjected to different power levels of infrared radiation, then we can use *one-way ANOVA*. However, besides infrared power if we are also interested in different exposure times, then *two-way ANOVA* needs to be employed.

6.1. One-Way ANOVA

There are 3 essential assumptions for the test to be accurate (Anon 2024)[7]:

1. Each group comes from a normal population distribution.
2. The population distributions have the same standard deviations ($\sigma_1=\sigma_2=\ldots=\sigma_n$).

   It is reasonable to expect that standard deviations of populations have some differences in values. Therefore, Peck *et al*. (2016) suggest that if $\sigma_{max} \leq 2 \cdot \sigma_{min}$ ANOVA still can safely be used.
3. The data are independent.

A similarity comparison of two-sample t-test and ANOVA is given by Moore *et al*. (2009). Suppose we are analyzing whether the means of two different groups of same size are different. Then we would employ two-sample t-test with equal variances (due to assumption #2):

$$t=\frac{\bar{X}-\bar{Y}}{S_p\sqrt{\frac{1}{n}+\frac{1}{n}}}=\frac{\sqrt{\frac{n}{2}}(\bar{X}-\bar{Y})}{S_p} \tag{6.1}$$

The square of test statistic is:

$$t^2=\frac{\frac{n}{2}(\bar{X}-\bar{Y})^2}{S_p^2} \tag{6.2}$$

7 https://online.stat.psu.edu/stat500/lesson/10/10.2/10.2.1

If we had used ANOVA, the *F-statistic* would have been exactly equal to t^2 computed using Eq. (6.2). A careful inspection of Eq. (6.2) reveals couple things:

1. The numerator measures the variation *between* the groups (known as fit).
2. The denominator measures the variation *within* groups (known as residual), see Eq. (4.3).

The null- and alternative-hypothesis for ANOVA are:

$$\begin{aligned} H_0&: \quad \mu_1=\mu_2=...=\mu_n \\ H_a&: \quad \text{At least two of the } \mu\text{'s are different} \end{aligned} \tag{6.3}$$

Therefore the basic idea is, to test H_0, we simply compare the variation *between* the means of the groups with the variation *within* groups. A graphical example adapted from Peck *et al*. (2016) can cement our understanding:

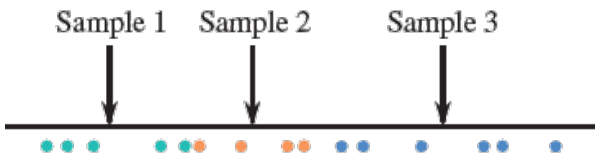


Fig 6.1-A: *A dataset with small within variability*

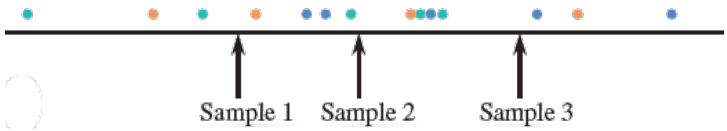


Fig 6.1-B: *A dataset with high within variability*

It is clearly seen from Fig. (6.1-A) that H_0 can be rejected as the means of 3 samples are different. The variability within each sample is smaller than the differences between the sample means.

In Fig. (6.1-B), the difference between sample means are as same as Fig. (6.1-A); however, there is considerable overlap between the samples. Therefore, *the difference between the means of the samples could simply be due to variability in sampling rather than the differences in population means.*

Computing the statistics:

Let k be the number of populations being compared [in Fig. (6.1) k=3] and $n_1, n_2, \ldots, n_k$ be the sample sizes:

1. Total number of observations:

   $N = n_1 + n_2 + \ldots + n_k$

2. Grand total (the sum of all observations):

$$T=\sum_{k=1}^{k}\sum_{i=1}^{n}X_{k,i}$$

3. Grand mean (average of all observations):

$$\bar{x}=\frac{T}{N}$$

4. Sum of squares of treatment:

$$SS_{TR}=n_1\cdot(\bar{x}_1-\bar{x})^2+n_2\cdot(\bar{x}_2-\bar{x})^2+\ldots+n_k\cdot(\bar{x}_k-\bar{x})^2$$

where *df=k-1*

5. Sum of squares of error:

$$SS_{Error}=(n_1-1)\cdot s_1^2+(n_2-1)\cdot s_2^2+\ldots+(n_k-1)\cdot s_k^2$$

where *df = N-k*

6. Mean squares:

$$MS_{TR}=\frac{SS_{TR}}{k-1}\quad\text{and}\quad MS_{Error}=\frac{SS_{Error}}{N-k}$$

The test statistics:

$$F=\frac{MS_{TR}}{MS_{Error}} \tag{**6.4**}$$

with $df_1=k-1$ and $df_2=N-k$.

Before proceeding with an example on ANOVA, let's further investigate Eq. (6.4). Remember that *F* distribution is the ratio of independent chi-square random variables and is given with the following equation:

$$F=\frac{U/m}{V/n} \tag{**6.5**}$$

where *U* and *V* are independent chi-square random variables with *m* and *n* degrees of freedom.

The following theorem establishes the link between Eqs. (6.4 & 6.5):

Theorem: Let $Y_1, Y_2, \ldots, Y_n$ be random sample from a normal distribution with mean μ and variance σ^2. Then,

$$\frac{(n-1)S^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \qquad \textbf{(6.6)}$$

has a chi-square distribution with *n-1* degrees of freedom. A proof of Eq. (6.6) is given by Larsen & Marx (2011) and is beyond the scope of this study.

Using Eq. (6.6), now it is easy to see that when sum of squares of treatment (or error) is divided by σ, it will have a chi-square distribution. Therefore Eq. (6.4) is indeed equivalent to Eq. (6.5) and therefore gives an F distribution with $df_1 = k-1$ and $df_2 = N-k$.

Example 6.1

In most of the integrated circuit manufacturing, a plasma etching process is widely used to remove unwanted material from the wafers which are coated with a layer of material, such as silicon dioxide. A process engineer is interested in investigating the relationship between the radio frequency power and the etch rate. The etch rate data (in Å/min) from a plasma etching experiment is given below:

160 W	180 W	200 W	220 W
575	565	600	725
542	593	651	700
530	590	610	715
539	579	637	685
570	610	629	710

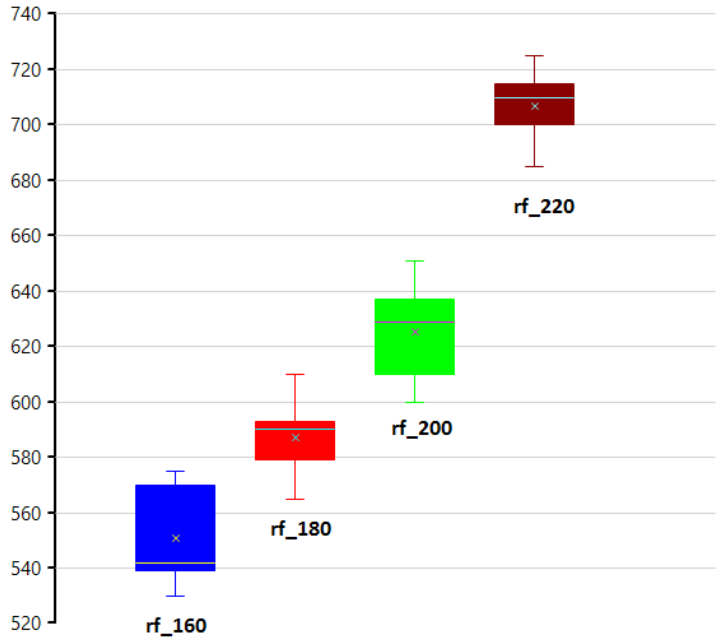
Does the RF power affect etching rate? (Adapted from Montgomery 2012)

Solution:

Before attempting any numerical solution, let's first visualize the data using box-whisker plot generated with a Python script:

Script 6.1

```
import scisuit.plot as plt

rf_160 = [575, 542, 530, 539, 570]
rf_180 = [565, 593, 590, 579, 610]
rf_200 = [600, 651, 610, 637, 629]
rf_220 = [725, 700, 715, 685, 710]

for dt in [rf_160, rf_180, rf_200, rf_220]:
    _name = [ k for k,v in locals().items() if v == dt][0]
    plt.boxplot(data=dt, label=_name)
plt.show()
```



It is immediately seen from the figure that μ_{220} is considerably different than other means. It can thus be inferred that the null hypothesis will be rejected since H_0 claims:

$$\mu_{160}=\mu_{180}=\mu_{200}=\mu_{220}$$

Fig 6.2: *The etch rate data at different RFs*

Before using *scisuit*'s built-in function, let's compute *F-value* using a Python script so that above-shown steps to calculate test statistics become clearer.

Script 6.2

```
import numpy as np
from scisuit.stats import qf

#create a 2D array
data = np.array([rf_160, rf_180, rf_200, rf_220]) #see Script (6.1)

#compute grand mean
grandmean = np.mean(data)

ss_tr, ss_error = 0, 0
for dt in data:
    n = len(dt) #size of each sample
    ss_tr += n*(np.mean(dt)-grandmean)**2
    ss_error += (n-1)*np.var(dt, ddof=1) #note ddof=1, the sample variance

row, col = data.shape
df_tr = row - 1
df_error = row*(col - 1)

Fvalue = (ss_tr/df_tr) / (ss_error/df_error)
Fcritical = qf(1-0.05, df1=df_tr, df2=df_error)

print(f"F={Fvalue}, F-critical={Fcritical}")
```

```
F=66.8, F-critical=3.24
```

Since the computed *F-value* is considerably greater than *F-critical*, we can safely reject H_0. Using *scisuit*'s built-in *aov* function:

Script 6.3

```
aovresult = aov(rf_160, rf_180, rf_200, rf_220)
print(aovresult)
```

One-Way ANOVA Results

Source	**df**	**SS**	**MS**	**F**	**p-value**
Treatment	3	66870.55	22290.18	66.80	2.8829e-09
Error	16	5339.20	333.70		
Total	19	72209.75			

Since $p<0.05$, we can reject H_0 in favor of H_1.

Now, had we not plotted Fig. (6.2), we would not be able to see why H_0 has been rejected. As a matter of fact, among other reasons due to overlap in whiskers and boxes or outliers a box-whisker plot does not always clearly show whether H_0 will be rejected. Therefore, we need to use post hoc tests along

with ANOVA. There are several tests[8] for this purpose, here we will be using Tukey's test[9]. Continuing from Script (6.3):

```
tukresult = tukey(alpha=0.05, aovresult=aovresult)
print(tukresult)
```

Tukey Test Results (alpha=0.05)

Pairwise Diff	**i-j**	**Interval**
1 - 2	-36.20	(-69.25, -3.15)
1 - 3	-74.20	(-107.25, -41.15)
1 - 4	-155.80	(-188.85, -122.75)
2 - 3	-38.00	(-71.05, -4.95)
2 - 4	-119.60	(-152.65, -86.55)
3 - 4	-81.60	(-114.65, -48.55)

Since none of the pairs contain the value 0.0, the Tukey procedure shows that means of all pairs are significantly different. Thus it can be concluded that each power level has an effect on etch rate that is different from the other power levels.

6.2. Two-Way ANOVA

In one-way ANOVA, the populations were classified according to a single factor; whereas in two-way ANOVA, as the name implies, there are two factors, each with different number of levels. For example, a baker might choose 3 different baking temperatures (150, 175, 200°C) and 2 different baking times (45 and 60 min) to optimize a cake recipe. In this example we have two factors (baking time and temperature) each with different number of levels (Devore *et al*. 2021; Moore *et al*. 2009).

Moore *et al*. (2009) lists the following advantages for using two-way ANOVA:

1. It is more efficient (i.e., less costly) to study two factors rather than each separately,
2. The variation in residuals can be decreased by the inclusion of a second factor,
3. Interactions between factors can be explored.

8 https://en.wikipedia.org/wiki/Post_hoc_analysis
9 https://en.wikipedia.org/wiki/Tukey%27s_range_test

In order to analyze a data set with two-way ANOVA the following assumptions must be satisfied (Field 2024; Moore 2012):

1. The response variable must be continuous (e.g., weight, height, yield, …),
2. The two independent variables must consist of discrete levels (e.g., type of treatment, brand of product) and each factor must have at least two levels,
3. In order to analyze interaction effects between independent variables, there should be replicates,
4. The observations must be independent,
5. It is desirable that the design should be balanced.

Let's start from **#5** and take a look at what it means balanced or unbalanced. In ANOVA or design of experiments, a balanced design has equal number of observations for all possible combinations of factor levels. For example[10], assume that the independent variables are A, B, C with 2 levels. Table (6.1) shows a balanced design whereas Table (6.2) shows an unbalanced design of the same factors (since the combination [1, 0, 0] is missing).

Table 6.1: *Balanced Design*

A	B	C
0	0	0
0	0	1
0	1	0
0	1	1
1	0	0
1	0	1
1	1	0
1	1	1

Table 6.2: *Unbalanced Design*

A	B	C
0	0	0
0	1	0
0	1	0
0	0	1
0	1	0
1	0	1
1	1	0
1	1	1

Note that if Table (6.1) was re-designed such that each row displayed a factor level (0 or 1) and each column displayed a factor (A, B or C) then there would be *no empty cells* in that table. If the data includes multiple observations for each treatment, the design includes *replication*.

10 https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/anova/supporting-topics/anova-models/balanced-and-unbalanced-designs/

Example 6.2

A study by Moore and Eddleman[11] (1991) investigated the removal of marks made by erasable pens on cotton and cotton/polyester fabrics. The following data compare three different pens and four different wash treatments with respect to their ability to remove marks on. The response variable is based on the color change and the lower the value the more marks were removed.

Table 6.3: *Effect of washing treatment and different pen brands on color change*

	Wash 1	**Wash 2**	**Wash 3**	**Wash 4**
Pen #1	0.97	0.48	0.48	0.46
Pen #2	0.77	0.14	0.22	0.25
Pen #3	0.67	0.39	0.57	0.19

Is there any difference in color change due either to different brands of pen or to the different washing treatments? (Adapted from *Devore et al*. 2021)

Solution:

The data satisfies the requirements to be analyzed with two-factor ANOVA, since:

1. There are two independent factors (pen brands and washing treatment),
2. The independent variables consist of discrete levels (e.g., brand #1, #2 and #3)
3. There are no empty cells (data is balanced),
4. There are ***no replicates*** (interaction cannot be explored),
5. Observations are independent.

Once a table similar to Table (6.3) is prepared, finding the F-values for both factors is fairly straightforward if a spreadsheet software is used.

Grand mean (T) = 0.466

11 **Moore MA, Eddleman VL** (1991). An Assessment of the Effects of Treatment, Time, and Heat on the Removal of Erasable Pen Marks from Cotton and Cotton/Polyester Blend Fabrics. *J. Test. Eval*.. 19(5): 394-397

Averages of treatments ($\mu_{treatments}$) = [0.803, 0.337, 0.423, 0.3]

$$SS_{treatment}=\sum_{i=1}^{4}(\mu_{treatments}[i]-T)^2\times 3=0.48 \quad \text{and} \quad MS_{treatment}=\frac{SS_{treatment}}{df}=\frac{0.48}{4-1}=0.16$$

Averages of brands (μ_{brands}) = [0.598, 0.345, 0.455]

$$SS_{brand}=\sum_{i=1}^{3}(\mu_{brands}[i]-T)^2\times 4=0.128 \quad \text{and} \quad MS_{brand}=\frac{SS_{brand}}{df}=\frac{0.128}{3-1}=0.06$$

$$SS_{Error}=\sum\sum(\mu_{ij}-T)-SS_{Treatment}-SS_{brand}=0.087 \quad \text{and} \quad MS_{Error}=\frac{SS_{Error}}{df}=\frac{0.087}{(3-1)\times(4-1)}=0.014$$

$$F_{treatment}=\frac{MS_{treatment}}{MS_{Error}}=\frac{0.16}{0.014}=11.05$$

$$F_{brand}=\frac{MS_{brand}}{MS_{Error}}=\frac{0.06}{0.014}=4.15$$

Although the solution is straightforward, it is still cumbersome and error-prone; therefore, it is best to use functions dedicated for this purpose:

Script 6.4

```
brand = [1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 3, 3]
treatment = [1, 2, 3, 4, 1, 2, 3, 4, 1, 2, 3, 4]
removal = [0.97, 0.48, 0.48, 0.46, 0.77, 0.14, 0.22, 0.25, 0.67, 0.39, 0.57, 0.19]

result = aov2(y=removal, x1=treatment, x2=brand)
print(result)
```

Two-way ANOVA Results

Source	df	SS	MS	F	p-value
x1	3	0.48	0.16	11.05	7.40e-03
x2	2	0.13	0.06	4.43	6.58e-02

■

Unlike Example (6.2) in which the data does not have replicates, the following example will demonstrate a data set which have replicates. It should be noted that when replicates are involved the solution becomes slightly more tedious and therefore the following example will be directly solved using *scisuit*'s built-in function. Interested readers can consult to textbooks (Devore *et al*. 2021) for a detailed solution.

Example 6.3

A process engineer is testing the effect of catalyst type (A, B, C) and reaction temperature (high, medium, low) on the yield of a chemical reaction. She designs an experiment with 3 replicates for each combination as shown in the following data. Do both catalyst type and reaction temperature have an effect on the reaction yield?

Catalyst = [A, A, A, A, A, A, A, A, A, B, B, B, B, B, B, B, B, B, C, C, C, C, C, C, C, C, C]

Temperature = [L, L, L, M, M, M, H, H, H, L, L, L, M, M, M, H, H, H, L, L, L, M, M, M, H, H, H]

%Yield = [85, 88, 90, 80, 82, 84, 75, 78, 77, 90, 92, 91, 85, 87, 89, 80, 83, 82, 88, 90, 91, 84, 86, 85, 79, 80, 81]

Solution:

If one wishes to use a spreadsheet for the solution, a table of averages needs to be prepared as shown below:

Table 6.4: *Effect of temperature and catalyst type on reaction rate*

	Temperature		
Catalyst	**L**	**M**	**H**
A	87.667	82	76.667
B	91	87	81.667
C	89.667	85	80

After preparing the above-shown table, a methodology similar to Example (6.2) can be followed.

Let’s solve the question directly by using *scisuit*’s built-in function:

Script 6.5

```
from scisuit.stats import aov2

Catalyst = ["A", "A", "A", "A", "A", "A", "A", "A", "A",
            "B", "B", "B", "B", "B", "B", "B", "B", "B",
            "C", "C", "C", "C", "C", "C", "C", "C", "C"]

Temperature = ["L", "L", "L", "M", "M", "M", "H", "H", "H",
               "L", "L", "L", "M", "M", "M", "H", "H", "H",
               "L", "L", "L", "M", "M", "M", "H", "H", "H"]

Yield = [85, 88, 90, 80, 82, 84, 75, 78, 77, 90, 92, 91,
         85, 87, 89, 80, 83, 82, 88, 90, 91, 84, 86, 85, 79, 80, 81]

result = aov2(y=Yield, x1=Temperature, x2=Catalyst)
print(result)
```

Two-way ANOVA Results

Source	**df**	**SS**	**MS**	**F**	**p-value**
x1	2	450.30	225.15	83.27	7.9886e-10
x2	2	90.74	45.37	16.78	7.7004e-05
*x1*x2*	4	3.04	0.76	0.28	8.8654e-01

From the ANOVA results, it is seen that both temperature and catalyst have significant ($p<0.05$) effect on reaction yield.

7. Linear Regression

Based on the amount of error associated with data, there are two general approaches for *curve fitting* (Chapra & Canale 2013):

1. *Regression:* When data shows a significant degree of error or "noise" (generally originates from experimental measurements), we want a curve that represents the *general trend* of the data.
2. *Interpolation:* When the noise in data can be ignored (generally originates from tables), we would like a curve(s) that pass directly through each of the data points.

In terms of mathematical expressions, interpolation (Eq. 7.1) and regression (Eq. 7.2) can be shown as follows:

$$Y = f(X) \tag{7.1}$$

$$Y = f(X) + \epsilon \tag{7.2}$$

Peck *et al*. (2016) used the terms *deterministic* and *probabilistic* relationships for Eq. (7.1) and Eq. (7.2), respectively. Therefore a *probabilistic relationship is actually a deterministic relationship with noise* (random deviations).

To further our understanding on Eq. (7.2), a simple example from Larsen & Marx (2011) can be helpful: Consider a tooling process where the initial weight of the sample determines the finished weight of the steel rods. For example, in a simple experiment if the initial weight was measured as 2.745 g then the finished weight was measured as 2.080 g. However, even if the initial weight is controlled and is exactly 2.745 g, in reality the finished weight would fluctuate around 2.080 g. and therefore, with each x (independent variable) there will be a range of possible y values (dependent variable), which Eq. (7.2) exactly tells us.

7.1. Simple Linear Regression

When there is only a single explanatory (independent) variable, the model is referred to as "simple" linear regression. Therefore, Eq. (7.2) can be expressed as:

$$Y = \beta_0 + \beta_1 x + \epsilon \tag{7.3}$$

where regardless of the x value, the random variable ε is assumed to follow a $N(0, \sigma)$ distribution.

Let x^* show a particular value of x, then:

$$E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^* = \mu_{Y|x^*} \tag{7.4}$$

$$Var(\beta_0 + \beta_1 x^* + \epsilon) = Var(\epsilon) = \sigma^2_{Y|x^*} \tag{7.5}$$

where the notation $Y|x^*$ should be read as the value of Y when $x=x^*$, i.e., the mean value of Y when $x=x^*$. Note also that Eq. (7.4) tells us something important that the population regression line is the line of mean values of Y.

The following assumptions are made for a linear model (Larsen & Marx, 2011):

1. $f_{Y|x}(y)$ is a normal probability density function for all x (i.e., for a known x value, there is a probability density function associated with y values)
2. The standard deviations, σ, of y-values are same for all x values.
3. For all x-values, the distributions associated with $f_{Y|x}(y)$ are independent.

Example 7.1

Suppose that the relationship between applied stress (x) and time to fracture (y) is given by the simple linear regression model with $\beta_0=65$, $\beta_1=-1.2$, and $\sigma=8$. What is the probability of getting a fracture value greater than 50 when the applied stress in 20? (Adapted from Devore et al. 2021)

Solution:

Let's compute y when x=20:

$$y = 65 - 1.2x = 65 - 1.2 \times 20 = 41$$

Note that if this was a curve fitting problem in nature, then whenever the stress value was 20, the fracture time would have always been equal to 41. However, since Eq. (7.2) tells us that random deviations are involved, this cannot be the case. We already know that the random deviations, namely ε, follows a normal distribution. Therefore, it becomes straightforward to compute the probability:

$$P\left(Z>\frac{50-41}{8}\right)=P\left(Z>1.125\right)=1-pnorm\left(1.125\right)=0.13 \qquad \blacksquare$$

In Example (7.1), the coefficients, namely β_0 and β_1, of the regression line was given. However, in practice we need to estimate these coefficients. It should be noted that there are two commonly[12] used methods for estimating the regression coefficients (please note that we use the word, *estimate*):

1. Least squares estimation method,
2. Maximum likelihood estimation method.

7.1.1. Least Squares Estimation

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ represent n observation pairs, from the measurement of X and Y. Our goal is to find β_0 and β_1 in Eq. (7.3) such that the drawn line is *as close as possible* to all data points.

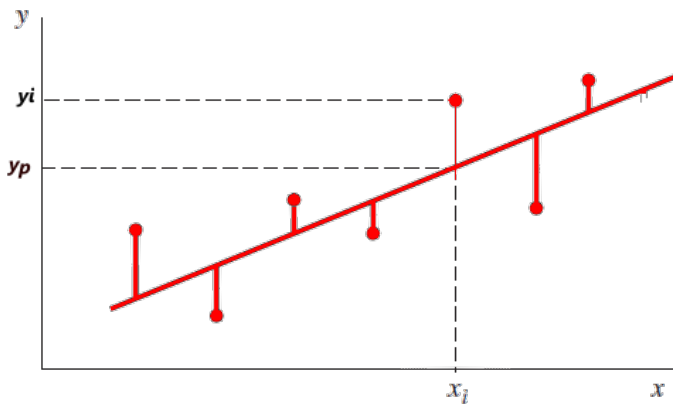


Fig 7.1: *Fitting a line through a set of data points*

In the figure y_i is the measured data point whereas y_p is the predicted value both of which corresponds to x_i. The associated error (also known as *residual*) is with this prediction is:

$$e_i = y_i - y_p$$

Since by definition we want the line as close as possible to all data points, therefore our goal is to minimize the sum of e_i's by varying β_0 and β_1.

12 https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/reliability/supporting-topics/estimation-methods/least-squares-and-maximum-likelihood-estimation-methods/

The residual sum of squares (RSS) also known as sum of squares of error (SSE):

$$RSS=\sum_{i=1}^{n} e_i^2=e_1^2+e_2^2+...+e_n^2 \tag{7.6}$$

If the coefficients of the best line passing through the data points are β_0 and β_1 then:

$$L=RSS=\sum_{i=1}^{n} (y_i-\beta_0-\beta_1 x)^2 \tag{7.7}$$

The partial derivatives of Eq. (7.7) with respect to β_0 and β_1 are:

$$\frac{\partial L}{\partial \beta_0}=\sum_{i=1}^{n} -2(y_i-\beta_0-\beta_1 \cdot x_i)=0$$

$$\frac{\partial L}{\partial \beta_1}=\sum_{i=1}^{n} 2x_i(y_i-\beta_0-\beta_1 \cdot x_i)=0$$

Dropping the constants -2 and 2 from both equations and simply rearranging the terms yields:

$$\sum_{i=1}^{n} y_i=n\beta_0+\beta_1\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{n} x_i \cdot y_i=\beta_0\sum_{i=1}^{n} x_i+\beta_1\sum_{i=1}^{n} x_i^2$$

We have two equations and two unknowns, therefore it is possible to solve this system of equations. Here, one can use the elimination method; however, Cramer's rule provides a direct solution. Let's solve for β_1 and leave β_0 as an exercise:

$$\hat{\beta}_1=\frac{\begin{vmatrix} n & \sum y_i \\ \sum x_i & \sum x_i y_i \end{vmatrix}}{\begin{vmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{vmatrix}}$$

If one takes the determinants in numerator and denominator, then:

$$\hat{\beta}_1 = \frac{n\sum x_i y_i - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \tag{7.8}$$

β_1 can be further simplified if a notation S_{xy} and S_{xx} and S_{xy} are defined as:

$$S_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n}(\sum x_i)^2$$

$$S_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}(\sum y_i)^2$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n}(\sum x_i)(\sum y_i)$$

Then $\hat{\beta}_1$ can be simplified as:

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} \tag{7.9}$$

and $\hat{\beta}_0$ is equal to:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \cdot \bar{x} \tag{7.10}$$

and the estimated variance is:

$$\hat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{7.11}$$

where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_i, \quad i = 1, 2, \ldots, n$

7.1.2. Maximum likelihood estimation

Before proceeding with the derivation based on maximum likelihood estimation (MLE), let's work on a simple example.

Example 7.2

Suppose you have been tasked with finding the probability of heads (H) and tails (T) of an unknown coin. You flipped the coin for 3 times and the sequence is HTH. What is the probability, *p*? (Adapted from Larsen & Marx)

Solution:

It makes sense with defining a random variable, X, as follows:

$$X=\begin{cases}1 & \text{heads come up}\\ 0 & \text{tails come up}\end{cases}$$

Then a probability model is defined:

$$p_X(k)=p^k(1-p)^{1-k}=\begin{cases}p & \text{k=1}\\ 1-p & \text{k=0}\end{cases}$$

Therefore, based on the probability model the function is that defines the sequence HTH is:

$$p_X(k)=p^2(1-p)$$

Using calculus, it can easily be computed that the value that maximizes the probability model is: $p=2/3$. ■

Now, instead of the sequence HTH (Example 7.2) we have data pairs $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ obtained from a random experiment. Furthermore, it is known that the y_i's are normally distributed with mean $\beta_0+\beta_1 x_i$ and variance σ^2 (Eqs. 7.4 & 7.5).

The equation for normal distribution is:

$$f_Z(z)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-1}{2}\cdot\left(\frac{x-\mu}{\sigma}\right)^2},-\infty<x<\infty \qquad \textbf{(7.12)}$$

Replacing x and μ in Eq. (7.12) with y_i and Eq. (7.4), respectively, yields the probability model for a single data pair:

$$f_Z(z)=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-1}{2}\cdot\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2} \quad \textbf{(7.13)}$$

For n data pairs, the maximum likelihood function is:

$$L=\prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{-1}{2}\cdot\left(\frac{y_i-\beta_0-\beta_1 x_i}{\sigma}\right)^2} \quad \textbf{(7.14)}$$

In order to find MLE of β_0 and β_1 partial derivatives with respect to β_0 and β_1 must be taken. However, Eq. (7.14) is not easy to work with as is. Therefore, as suggested by Larsen and Marx (2011), taking the logarithm will make it more convenient to work with.

$$-2lnL=n\cdot\ln(2\pi)+n\ln(\sigma^2)+\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\beta_0-\beta_1\cdot x_i)^2 \quad \textbf{(7.15)}$$

Taking the partial derivatives of Eq. (7.15) with respect to β_0 and β_1 and solving the resulting set of equations similar to as shown in section (7.1.1) will yield Eqs. (7.9 & 7.10).

7.1.3. Properties of Linear Estimators

Due to the assumptions made for a linear model (section 7.1), the estimators, $\hat{\beta}_0, \hat{\beta}_1$ and $\hat{\sigma}$, are random variables (i.e., probability distribution functions are associated with them). Then,

1. $\hat{\beta}_0$ and $\hat{\beta}_1$ are normally distributed.
2. $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased, therefore, $E(\hat{\beta}_0)=\beta_0$ and $E(\hat{\beta}_1)=\beta_1$
3. $Var(\hat{\beta}_1)=\dfrac{\sigma^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}$
4. $Var(\hat{\beta}_0)=\dfrac{\sigma^2\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\bar{x})^2}$

Proof of #2:

In section (2.1.1), it was mentioned that to be an unbiased estimator, $E(\Theta) = \theta$ must be satisfied. In the case of $\hat{\beta}_1$, we need to show that $E(\hat{\beta}_1)=\beta_1$. If Eq. (7.8) is divided by n, the following equation is obtained:

$$\hat{\beta}_1=\frac{\sum x_i y_i-\frac{1}{n}(\sum x_i)(\sum y_i)}{\sum x_i^2-\frac{1}{n}(\sum x_i)^2} \quad \textbf{(I)}$$

Noting that $\bar{x}=\frac{\sum x_i}{n}$ the Eq. (I) can be rewritten as:

$$\hat{\beta}_1=\frac{\sum x_i y_i-\bar{x}\sum y_i}{\sum x_i^2-n\bar{x}^2} \quad \textbf{(II)}$$

Rearranging the terms in the numerator:

$$\hat{\beta}_1=\frac{\sum y_i(x_i-\bar{x})}{\sum x_i^2-n\bar{x}^2} \quad \textbf{(III)}$$

Note that due to the assumption of the linear model, in Eq. (III) except y_i, the other terms can be treated as constant. Therefore, replacing the expected value of y_i with Eq. (7.4) gives:

$$E(\hat{\beta}_1)=\frac{\sum(\beta_0+\beta_1 x_i)(x_i-\bar{x})}{\sum x_i^2-n\bar{x}^2} \quad \textbf{(IV)}$$

Expanding the terms in the numerator:

$$E(\hat{\beta}_1)=\frac{\beta_0\sum(x_i-\bar{x})+\beta_1\sum(x_i-\bar{x})x_i}{\sum x_i^2-n\bar{x}^2} \quad \textbf{(V)}$$

Noting that the first term in the numerator equals to 0 and the remaining terms in the numerator (except β_1) equals to the denominator, the proof is completed.

$$E(\hat{\beta}_1)=\beta_1 \quad \textbf{(V)}$$

A similar proof can be obtained for β_0. For cases *#3* and *#4*, Larsen & Marx (2011) presented a detailed proof.

Example 7.3

It seems logical that riskier investments might offer higher returns. A study by Statman *et al*. (2008)[13] explored this by conducting an experiment. One group of investors rated the risk (**x**) of a company's stock on a scale from 1 to 10, while a different group rated the expected return (**y**) on the same scale. This was done for 210 companies, and the average risk and return scores were calculated for each. Data for a sample of ten companies, ordered by risk level, is given below:

x = [4.3, 4.6, 5.2, 5.3, 5.5, 5.7, 6.1, 6.3, 6.8, 7.5]

y = [7.7, 5.2, 7.9, 5.8, 7.2, 7, 5.3, 6.8, 6.6, 4.7]

How does the risk of an investment related to its expected return? (Adapted from Devore et al. 2021)

Solution:

Let's first visualize the data using a scatter plot.

Script 7.1

```
import scisuit.plot as plt

x = [4.3, 4.6, 5.2, 5.3, 5.5, 5.7, 6.1, 6.3, 6.8, 7.5]
y = [7.7, 5.2, 7.9, 5.8, 7.2, 7, 5.3, 6.8, 6.6, 4.7]
plt.scatter(x=x, y=y)
plt.show()
```

13 **Statman M, Fisher KL, Anginer D** (2008). Affect in a Behavioral Asset-Pricing Model. Financial Analysts Journal, 64-2, 20-29.

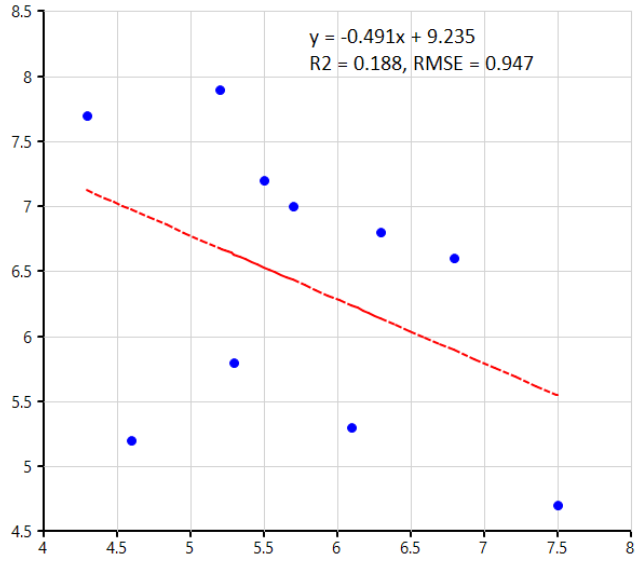


Fig 7.2: *Relationship between risk and expected return*

It is seen that there is a weak inverse relationship between the perceived risk of a company's stock and its expected return value.

Note: After plotting the data, since **scisuit**'s charts are interactive, the trendline was added by first selecting the data and then selecting "Add trendline" option.

Fig. (7.2) shows that there is no convincing relationship between risk and expected return of an investment. Let's take a look if this is numerically the case. Continuing from Script (7.1):

Script 7.2

```
from scisuit.stats import linregress
result = linregress(yobs=y, factor=x)
print(result)
```

```
Simple Linear Regression
F=1.85, p-value=0.211, R2=0.19

The regression equation: Y = 9.235 - 0.491·X
```

Predictor	**Coeff**	**StdError**	**T**	**p-value**
Intercept	9.235	2.10	4.40	0.0023
Slope	-0.491	0.36	-1.36	0.2110

Since $p>0.05$, we cannot reject the null hypothesis ($H_0: \beta_1=0$) in favor of H_1.

Have we carried out a reliable analysis, i.e., is there no relationship between risk and expected returns? Devore *et al*. (2021) suggested that with small number of observations, it is possible not to detect a relationship because when the sample size is small hypothesis tests do not have much power. Also note that the original study uses 210 observations where Statman *et al*. (2008) concluded that risk is a useful predictor of expected return, although the risk only accounted for 19% of expected returns. ■

7.2. Multiple Linear Regression

Suppose the taste of a fruit juice is related to sugar content and pH. We wish to establish an empirical model, which can be described as follows:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{7.16}$$

where y is the response variable (taste) and x_1 and x_2 are independent variables (sugar content and pH). Unlike simple linear regression (SLR) model, where only one independent variable exists, in multiple linear regression (MLR) problems at least 2 independent variables are of interest to us. Therefore, in general, the response variable maybe related to k independent (regressor) variables. The model is:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \epsilon \tag{7.17}$$

This model describes a hyperplane and the regression coefficient, β_j, represents the expected change in response to per unit change in x_j when all other variables are held constant (Montgomery 2012). If one enters the data in a spreadsheet, it would generally be in the following format:

Table 7.1: *Data for multiple linear regression*

y	**x_1**	**x_2**	**…**	**x_k**
y_1	x_{11}	x_{12}	…	x_{1k}
y_2	x_{21}	x_{22}	…	x_{2k}
y_n	x_{n1}	x_{n2}	…	x_{nk}

y is the response variable and *x* are the regressor variables. It is assumed that $n>k$.

The model equation for the data in Table (7.1):

$$y = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} + \epsilon_i, \quad i = 1, 2, \ldots, n \tag{7.18}$$

For example, for the 1st row (i=1) in Table (7.1), Eq. (7.18) yields, $y_1 = \beta_0 + \beta_1 \cdot x_{11} + \beta_2 \cdot x_{12} + \ldots + \beta_k \cdot x_{1k}$.

To find the regression coefficients, we will use a similar approach presented in section (7.1.1), such that the sum of the squares of errors, ε_i, is minimized. Therefore,

$$L=\sum_{i=1}^{n}\left(y_i-\beta_0-\sum_{j=1}^{k}\beta_j x_{ij}\right)^2 \tag{7.19}$$

where the function L will be minimized with respect to β_0, β_1, …, β_k which then will give the least square estimators, $\hat{\beta}_1,\hat{\beta}_2,..,\hat{\beta}_k$. The derivatives with respect to β_0 and β_j are:

$$\left.\frac{\partial L}{\partial \beta_0}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k}=-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right) \tag{7.20-a}$$

$$\left.\frac{\partial L}{\partial \beta_j}\right|_{\hat{\beta}_0,\hat{\beta}_1,\ldots,\hat{\beta}_k}=-2\sum_{i=1}^{n}\left(y_i-\hat{\beta}_0-\sum_{j=1}^{k}\hat{\beta}_j x_{ij}\right)x_{ij} \tag{7.20-b}$$

After some algebraic manipulation, Eq. (7.20) can be written in matrix notation as follows:

$$\begin{bmatrix} n & \sum x_{i1} & \sum x_{i2} & \ldots & \sum x_{ik} \\ \sum x_{i1} & \sum x_{i1}^2 & \sum x_{i1}x_{i2} & \ldots & \sum x_{i1}x_{ik} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ \sum x_{ik} & \sum x_{ik}x_{i1} & \sum x_{ik}x_{i2} & \ldots & \sum x_{ik}^2 \end{bmatrix}\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\\ \vdots\\ \hat{\beta}_k\end{bmatrix}\begin{bmatrix}\sum y_i\\ \sum x_{i1}y_i\\ \vdots\\ \sum x_{ik}yi\end{bmatrix} \tag{7.21}$$

which can be condensed to the following expression:

$$X\cdot\beta=y \tag{7.22}$$

Note that since **X** is an *i* by *k* matrix, therefore not square, the inverse does not exist and therefore the equation cannot be solved. The least-squares approach to solving Eq. (7.22) is by multiplying with transpose of **X**:

$$X^T X\cdot\beta=X^T\cdot y \tag{7.23}$$

The test of significance of regression involves the hypotheses:

$$\begin{aligned} &\boldsymbol{H_0}:\beta_1=\beta_2=...=\beta_k=0 \\ &\boldsymbol{H_1}:\beta_j\neq 0 \qquad \text{for at least for 1 j} \end{aligned} \tag{7.24}$$

Example 7.4

A process engineer who was tasked to improve the viscosity of a polymer, among the several factors, chose two process variables: reaction temperature and feed rate. She ran 16 experiments and collected the following data:

Temperature = [80, 93, 100, 82, 90, 99, 81, 96, 94, 93, 97, 95, 100, 85, 86, 87]

Feed Rate = [8, 9, 10, 12, 11, 8, 8, 10, 12, 11, 13, 11, 8, 12, 9, 12]

Viscosity = [2256, 2340, 2426, 2293, 2330, 2368, 2250, 2409, 2364, 2379, 2440, 2364, 2404, 2317, 2309, 2328]

Explain the effect of feed rate and temperature on polymer viscosity. (Adapted from Montgomery 2012).

Solution:

The solution involves several computations which can be performed by using a spreadsheet or by using *Python* with *numpy* library. Step by step solution for the coefficients can be found in the textbook from Montgomery (2012). We will be skipping all these steps and directly solve it using *scisuit*'s builtin *linregress* function.

Script 7.3

```
from scisuit.stats import linregress

#input values
temperature = [80, 93, 100, 82, 90, 99, 81, 96, 94, 93, 97, 95, 100, 85, 86, 87]
feedrate = [8, 9, 10, 12, 11, 8, 8, 10, 12, 11, 13, 11, 8, 12, 9, 12]
viscosity = [2256, 2340, 2426, 2293, 2330, 2368, 2250, 2409, 2364, 2379, 2440, 2364, 2404, 2317, 2309, 2328]

#note the order of input to factor
result = linregress(yobs=viscosity, factor=[temperature, feedrate])
print(result)
```

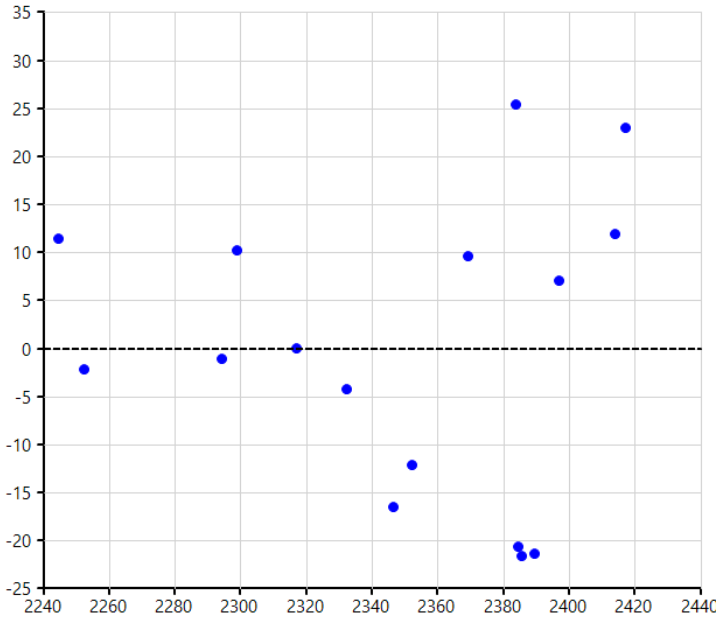
```
Multiple Linear Regression
F=82.5, p-value=4.0997e-08, R2=0.93

Predictor      Coeff      StdError      T          p-value
X0             1566.078   61.59         25.43      9.504e-14
X1             7.621      0.62          12.32      3.002e-09
X2             8.585      2.44          3.52       3.092e-03
```

Based on Eq. (7.24), the p-value tells us that at least one of the two variables (temperature and feed rate) has a nonzero regression coefficient. Furthermore, analysis on individual regression coefficients show that both temperature and feed rate have an effect on polymer's viscosity.

According to Larsen & Marx (2011), applied statisticians find residual plots to be very helpful in assessing the appropriateness of fitting. Continuing from Script (7.3), let's plot the residuals:

Script 7.4

```
import scisuit.plot as plt
import scisuit.plot.gdi as gdi

#x=Fits, y=Residuals
plt.scatter(x=result.Fits, y= result.Residuals)

#show a line at y=0
x0, x1 = min(result.Fits), max(result.Fits)*1.005
gdi.line(p1=(x0,0), p2=(x1, 0), lw=2, ls = "---" )

plt.show()
```



It is seen that the magnitudes of the residuals are comparable and they are randomly distributed. Therefore, the applied regression can be considered as appropriate.

Fig 7.3: *Fits vs residuals (y-axis)*

8. References

Box GEP., Hunter WG, Hunter JS (2005). Statistics for Experimenters: Design, Innovation, and Discovery, 2nd Ed., Wiley.

Carlton MA, Devore JL (2014). Probability with Applications in Engineering, Science and Technology. Springer USA.

Chapra SC, Canale RP (2013). Numerical methods for engineers, seventh edition. McGraw Hill Education.

Devore JL, Berk KN, Carlton MA (2021). Modern Mathematical Statistics with Applications. 3rd Ed., Springer.

Forbes C, Evans M, Hastings N, Peacock B (2011). Statistical Distributions, 4th Ed., Wiley.

Hastie T, Tibshirani R, Friedman J (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hogg RV, McKean JW, Craig AT (2019). Introduction to mathematical statistics, 8th Ed., Pearson.

Kanji GK (2006). 100 Statistical Tests, 3rd Ed., Sage Publications.

Kreyszig E, Kreyszig H, Norminton EJ (2011). Advanced Engineering Mathematics, 10th Ed., John Wiley & Sons Inc.

Larsen RJ, Marx ML (2011). An Introduction to Mathematical Statistics and Its Applications. 5th Ed., Prentice Hall.

Miller I, Miller M (2014). John E. Freund's Mathematical Statistics with Applications. 8th Ed., Person New International Edition.

Montgomery DC (2012). Design and analysis of experiments, 8th Ed., John Wiley & Sons, Inc.

Montgomery DC, Peck EA, Vining GG (2021). Introduction to Linear Regression Analysis, 6th Ed., Wiley.

Moore DS, McCabe GP, Craig BA (2009). Introduction to the Practice of Statistics. 6th Ed., W. H. Freeman and Company, New York.

Peck R, Olsen C, Devore JL (2016). Introduction to Statistics and Data Analysis. 5th Ed., Cengage Learning.

Pinheiro, CAR, Patetta M (2021). Introduction to Statistical and Machine Learning Methods for Data Science. Cary, NC: SAS Institute Inc.

Wackerly DD, Mendenhall W, Scheaffer RL (2008). Mathematical Statistics with Applications, 7th Ed., Thomson/Brooks Cole.