

MATLAB Workbook

CME106

**Introduction to Probability and
Statistics for Engineers**

First Edition

Vadim Khayms

Table of Contents

- 1. Random Number Generation**
- 2. Probability Distributions**
- 3. Parameter Estimation**
- 4. Hypothesis Testing (single population)**
- 5. Hypothesis Testing (two populations)**
- 6. Regression and Correlation Analyses**

Probability & Statistics

1. Random Number Generation

Commands:	unidrnd	Generates a discrete random number or a random vector
	unifrnd	Generates a random number or a vector from a uniform distribution
	hist	Creates a frequency plot (histogram)

We will attempt to solve one of the homework problems numerically by performing a virtual experiment using MATLAB's random number generator. The problem is as follows. Two points a and b are selected at random along the x -axis such that $-2 \leq b \leq 0$ and $0 \leq a \leq 3$. Find the probability that the distance between a and b is greater than 3 by performing one million trials. Make a histogram of the generated distances between x and y .

SOLUTION

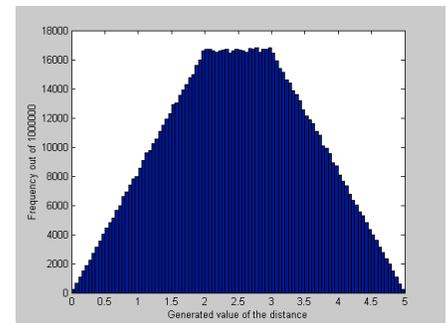
Two vectors of random numbers each with one million elements will be drawn from a uniform distribution and the distance between the corresponding elements compared to 3. See code below.

```
clear;
x=unifrnd(-2,0,1,1000000);
y=unifrnd(0,3,1,1000000);
s=0;
d=abs(x-y);
for i=1:1000000
    if d(i)>3
        s=s+1;
    else
        end
end
hist(d,100)
Probability=s/1000000
```

Running the code:

Probability =

0.3345



% The exact analytical solution is 1/3

Hint: generating a vector of random numbers all at once is computationally more efficient than generating random values one at a time within a for loop !

YOUR TURN

Using MATLAB, verify your answer to one other homework problem, which follows. A point is selected at random inside a circle. Find the probability that the point is closer to the center of the circle than to its circumference.

Hint: generate a random vector of x and y coordinates first, then consider only those points, which are inside the circle

2. Probability Distributions

Commands:	binopdf	PDF for the binominal distribution
	binocdf	CDF for the binominal distribution
	binoinv	Inverse CDF for the binomial distribution
	binornd	Generates random numbers from the binomial distribution
	poisspdf	PDF for the Poisson distribution
	poisscdf	CDF for the Poisson distribution
	poissinv	Inverse CDF for the Poisson distribution
	poissrnd	Generates random numbers from the Poisson distribution
	normpdf	PDF for the normal distribution
	normcdf	CDF for the normal distribution
	norminv	Inverse CDF for the normal distribution
	normrnd	Generates random numbers from the normal distribution

- a) A single lot containing 1000 chips manufactured by a semiconductor company is known to contain 1% of defective parts. What is the probability that at most 10 parts are defective in the lot? Use both the binominal and the Poisson distributions to obtain the answer.
- b) The diameters of shafts manufactured by an automotive company are normally distributed with the mean value of 1 inch and the standard deviation of 0.01 in. Each shaft is to be mounted into a bearing with an inner diameter of 1.025 in. Write a MATLAB script to estimate the proportion of defective parts out of 10,000, i.e. the fraction of all parts that do not fit into the bearings. On the same set of axes plot a histogram showing the observed frequencies of the shaft diameters and the scaled density function.

SOLUTION

Part a)

```
>> binocdf(10,1000,0.01)
ans =
    0.5830
```

```
>> poisscdf(10,10)
ans =
    0.5830
```

Part b)

```
clear
% x - observed number of defective
parts
% y - expected number of defective
parts
x=normrnd(1,0.01,1,10000);
s=0;
```

```
for i=1:10000
    if x(i)>1.025
        s=s+1;
    else end
end
Probability=s/10000
```

```
figure(2)

hist(x,100);

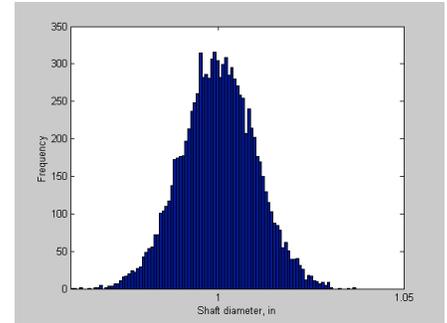
xlabel('Shaft diameter, in');

ylabel('Frequency');

Running the code:

Probability =

    0.0063
```



YOUR TURN

Write a script to simulate a sampling procedure from 1,000 identical lots using the binomial distribution with probability of a part being defective equal to 0.01. Create a histogram showing the frequencies of the number of observed defective parts in 1,000 lots using 10,000. On the same set of axes plot the scaled probability density function for the binomial distribution with the appropriate parameters. The density function can be plotted over the range from 0 to 25. Compare the empirical and the theoretical distributions. Note that the binomial distribution function **binopdf** returns non-zero values only when the argument is an integer.

3. Parameter Estimation

Commands:

normfit	Estimate parameters of a normal distribution
poissfit	Estimate parameters of a Poisson distribution
expfit	Estimate parameters of an exponential distribution

The time to failure of field-effect transistors (FETs) is known to be exponentially distributed. Data (in hours) has been collected for 10 samples. Using the data in the table below, compute the maximum likelihood estimate and the 95% confidence interval for the mean time to failure.

2540	2800
2650	2550
2420	2300
2580	2460
2470	2490

SOLUTION

```
>> x=[2540, 2650, 2420, 2580, 2470,
2800, 2550, 2300, 2460, 2490];
>> [mu_hat,
mu_confidence]=expfit(x,0.05)
mu_hat =
2526
mu_confidence =
1.0e+003 *
1.2113
4.3156
```

YOUR TURN

A traffic engineer would like to make an assessment of traffic flow at a busy intersection during weekday rush hour. The number of arrivals is believed to satisfy the conditions for the Poisson distribution. The engineer records the number of cars arriving at the intersection between the hours of 8am and 9am over a period of two weeks (10 business days). Using the data in the table (in cars/hr), compute the maximum likelihood estimate and the 95% confidence interval for the average number of car arrivals.

240	205
225	275
300	320
280	210
215	240

4. Hypothesis Testing (single population)

Commands:

ztest	z-test for the mean of a normal distribution with known variance
ttest	t-test for the mean of normal distribution with unknown variance

The wearout time of a part is known to be approximately 1000 hours. A new manufacturing process is proposed to increase the wearout time by over 200 hours. Data collected on 10 samples yields the following times in hours (see table below). The new process will be implemented at the factory only if it can be shown to result in an improvement in the wearout time of over 300 hours. Assume the wearout time to be normally distributed with the standard deviation of 20 hours. Should the proposed process be adopted ?

1350	1210
1320	1290
1290	1350
1280	1300
1340	1310

SOLUTION

```
>> x=[1350, 1320, 1290, 1280, 1340,
1210, 1290, 1350, 1300, 1310];
>> mean = 1300;
>> sigma = 20;
>> alpha = 0.05;
>> tail = 1;
>>
[hypothesis,p_value]=ztest(x,mean,sig
ma,alpha,tail)
```

```
hypothesis=
0
p_value =
0.2635
% since the p_value > 0.05, the null
hypothesis is accepted and the % new
process is not adopted
```

Hint: the *ztest* function can be used to perform right-sided, left-sided, and two-sided tests by specifying the “tail” parameter, which could be set to 1, 0, and -1. Type *help ztest* for details.

YOUR TURN

Repeat the above test, but now assuming that the standard deviation for the distribution of the wearout times is not known and must be estimated from the data.

5. Hypothesis Testing (two populations)

Commands:	tttest2	<i>t</i> -test for the difference of two means of normal distribution with unknown but equal variances
	tttest	<i>t</i> -test for the mean of normal distribution with unknown variance
	signtest	sign test for the mean of an arbitrary distribution

It is desired to compare the performance of two types of gasoline engines by comparing their fuel consumption for a certain number of miles traveled. No historical data is available from which to determine the variance of the fuel consumption, however, it is known that the two engines have similar characteristics and that any variability in the fuel usage should be

common to both types of engines. It is also assumed that the fuel consumption is approximately normally distributed. The fuel consumption data in gallons for three engines of the first type and 4 engines of the second type are shown in the table below. Can it be concluded that the two types of engines demonstrate different performance at the 10% significance level ?

Type 1, gallons	Type 2, gallons
540	575
520	540
535	560
---	545

SOLUTION

```
% This is a two-sided test, since the
question is to % compare two
populations without any prior
knowledge % which of the two means
is larger

>> x=[540, 520, 535];
>> y=[575, 540, 560, 545];

>>
[hypothesis,p_value]=ttest2(x,y,0.1)
hypothesis =
     1
p_value =
     0.0794

% since the p_value < 0.1, the null
hypothesis is rejected % and it is
concluded that the fuel consumptions
are not % the same at the 10%
significance level
```

```
% We can perform a "hand"
calculation to compare the % results
with the results of ttest2 function

>> s_squared=(3-1)/(3+4-
2)*var(x)+(4-1)/(3+4-2)*var(y)
s_squared =
    193.3333

>> test_statistic=(mean(x)-
mean(y))/sqrt(s_squared)/sqrt(1/3+1/4
)
test_statistic =
    -2.1972

>> p_value=2*tcdf(test_statistic,3+4-
2)
p_value =
     0.0794

% The p-value is in agreement to the
output from the % ttest2 function
```

YOUR TURN

- a) The test engineers have examined data on gasoline consumption and have noticed that the fuel consumption for the two types of engines has been determined pair-wise over different distances traveled. Fuel consumption is assumed to be normally distributed. Due to the inherent variability in the data associated with the various distances, it is believed that a paired *t*-test would be more appropriate. For the paired fuel consumption data provided in the table below, can it be concluded at the 10% significance level that the two engines consume fuel at different rates ?

Type 1, gallons	Type 2, gallons
540	555
520	515
580	585
500	505

- b) Suppose that because of the lack of historical data, test engineers are not certain that the fuel consumption in a) is normally distributed. Rather than using a paired t -test, they are considering using a distribution-free sign test. Using the data in the table above, what conclusion is reached at the 5% significance level ?

6. Regression and Correlation Analyses

Commands:	polyfit	estimates coefficients for a polynomial fit to a set of paired data
	lsline	superimposes a best-fit line to a set of scattered data
	corrcoef	sample correlation coefficient

Thermistors are passive devices frequently used to measure temperature. They are resistive elements whose resistance increases as a function of temperature. To calibrate a thermistor, a voltage is applied across its terminals and the value of the current is recorded for a given temperature. The data collected is shown in the table below (current in Amperes and voltage in Volts).

- a) For each value of the temperature plot on the same set of axes a linear curve fit of the voltage versus current through the thermistor and separately a calibration curve (resistance versus temperature)
- b) Compute the coefficient of determination for each of the five sets of data

T=10 C		T=20C		T=30C		T=40C		T=50C	
I	V	I	V	I	V	I	V	I	V
0.11	1	0.08	1	0.07	1	0.04	1	0.02	1
0.21	2	0.15	2	0.11	2	0.07	2	0.05	2
0.32	3	0.23	3	0.17	3	0.11	3	0.07	3
0.42	4	0.33	4	0.23	4	0.15	4	0.08	4

SOLUTION

```
% Input data
>> V=[1 2 3 4];
>> I10=[0.11 0.21 0.32 0.42];
>> I20=[0.08 0.15 0.23 0.33];
>> I30=[0.07 0.11 0.17 0.23];
>> I40=[0.04 0.07 0.11 0.15];
>> I50=[0.02 0.05 0.07 0.08];

% Compute regression coefficients
>> [p10,s10]=polyfit(I10,V,1);
>> [p20,s20]=polyfit(I20,V,1);
>> [p30,s30]=polyfit(I30,V,1);
>> [p40,s40]=polyfit(I40,V,1);
>> [p50,s50]=polyfit(I50,V,1);

% Plot scatter data and regression
lines
>> figure(1)
>> grid on
>> hold on
>> plot(I10,V,'.');
>> lsline
>> plot(I20,V,'.');
>> lsline
>> plot(I30,V,'.');
>> lsline
>> plot(I40,V,'.');
>> lsline
>> plot(I50,V,'.');
>> lsline
>> hold off
>> xlabel('Current, amperes');
>> ylabel('Voltage, volts');

% Plot calibration curve, i.e. slope of
regression lines % versus
temperature
>> figure(2)
```

```
>> T=[10 20 30 40 50];
>> resistance=[p10(1) p20(1) p30(1)
p40(1) p50(1)];
>> plot(T,resistance,'+')
>> axis([5,55,5,50]);
>> grid on
>> xlabel('Temperature, C');
>> ylabel('Resistance, Ohm');
```

% The coefficient of determination is equal to the square % of the sample correlation coefficient. The diagonal % elements are scaled variances (all equal to 1. The off-% diagonal elements give the correlation coefficient.

```
>> rho10=corrcoef(V,I_10)

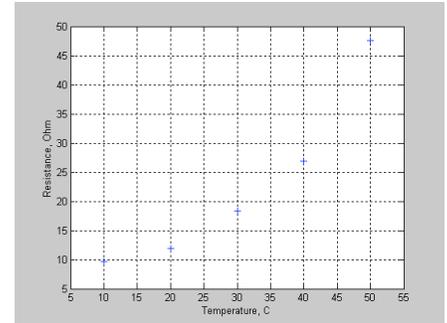
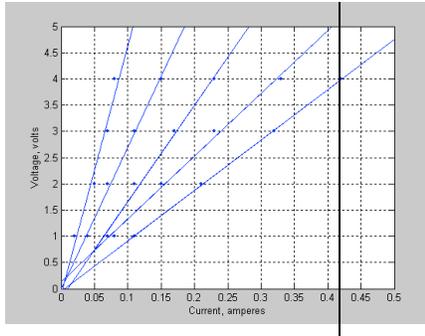
rho10 =
    1.0000    0.9998
    0.9998    1.0000
>> rho20=corrcoef(V,I_20)

rho20 =
    1.0000    0.9967
    0.9967    1.0000
>> rho30=corrcoef(V,I_30)

rho30 =
    1.0000    0.9959
    0.9959    1.0000
>> rho40=corrcoef(V,I_40)

rho40 =
    1.0000    0.9978
    0.9978    1.0000
>> rho50=corrcoef(V,I_50)

rho50 =
    1.0000    0.9759
    0.9759    1.0000
```



 **YOUR TURN**

The voltage-current characteristic of a newly developed non-linear electronic device is thought to have a quadratic dependence of the form: $I = c_0 + c_1V + c_2V^2$. A set of measurements was performed to estimate the regression coefficients c_0, c_1, c_2 . The measurements of the voltage and current are provided in the table below. Determine the regression coefficients. Plot on the same set of axes the current-voltage data and the quadratic curve fit.

V (volts)	I (Amps)
1	0.3
2	3.5
3	12
4	20
5	29
6	43
7	55
8	75