PGEG371: Data Analysis & Geostatistics

Selection Calculation and other PDF Distributions

Laboratory Exercise # 4

made by Dr. Sandra Vega and Mr. James Small

Student ID:

Name:

Read through this instruction sheet then answer the 'pre-Lab' quiz BEFORE starting the exercises!

1. Aim

The purpose of this laboratory exercise is to use the concept of cut-off in sample data and the application of grade/tonnage curves.

On successful completion of this exercise, you should be able to

- Produce selection calculations
- Graph those calculations as grade/tonnage curves

2. Introduction

So far, we have learned the relationships between sample data and population when the population follows a normal distribution. Sometimes we want to know what proportion of a population is in an interval, e.g. what proportion of the population is taller than 180 cm. Using a cut-off value we cut our population in two portions, the below cut-off and the above cut-off. Sometimes we do not know what the cut-off would be, and then we use a series of cut-offs and produce what miners call a "grade/tonnage" curve.

3. **Formulas used (Refer to** the height of the students – pages 46 and 47 of textbook).

a. ***Standardised Normal***, where x is the standardised normal height of g student, g_c is the cut-off value, μ^* is the estimated average of the population and σ^* is the estimated standard deviation of the population

$$x = \frac{g_c - \mu^*}{\sigma^*}$$

b. Formula to calculate the ***population above the cut-off***, where $\Phi(x)$ is the normal of the population that lies below the cut-off, obtained from Table 1.

$$P = 1 - \Phi(x), \quad x \geq 0$$
$$P = \Phi(|x|), \quad x < 0$$

c. Formula to calculate the Density probability (Standardized)

$$\phi(x) = \frac{1}{\sqrt{2\pi}} e^{\left\{\frac{-x^2}{2}\right\}}$$

d. Formula to calculate ***the average of the portion of our population who lie at or above the cut-off***
Where,
μ_c^* is estimate of average of students in the population that lies at or above the cut-off;

$$\mu_c^* = \mu^* + \frac{\sigma^*}{P} \phi(x)$$

μ^* and σ^* are the estimates for the mean and standard deviation of the complete population;
P is the proportion of the students population that lies at or above the cut-off and
$\varphi(x)$ is the height of the standard Normal graph at the cut-off.

MATLAB Functions

Some simple yet important statistical measures might include these in the following list. The equivalent MATLAB internal functions are shown in the right hand column.

Labelling X and Y axis	**xlabel('text')** adds text beside the X-axis on the current axis (same for **ylabel('text')**).
Mean for sample data and for population	**mean(x) –** best estimator of mean for a 'population' **mean(x,1) –** mean of the 'sample data'
Std for sample data and for population	**std(x) –** best estimator of std for a 'population' **std(x,1) –** std of the 'sample data'
normfit	**[mu,std,mu_conf,std_conf]=normfit(data)** provides the best estimate for mu, std, and confidence intervals for mu (mu_conf) and std (std_conf) of data.
Distribution Tool	**disttool** a distribution tool that displays an "ideal" normal distribution pdf and cdf.
Plotting two Ys and one X	**plotyy(cutoff,P,cutoff, mu*)** - Plotting of a cut-off vs. Probability and Mean.
square root of x	**sqrt(x)**
exponential function of x	**exp(x)**
Visual PDF and CDF	**ksdensity (x, 'function','pdf')** computes a probability density estimate of the sample in the vector x. **ksdensity (x, 'function','cdf')** computes a cumulative probability estimate of the sample in the vector x.

REMEMBER:

- *Open Matlab*
- *Set your current directory in your '**home directory**';*
- *Open a diary file with the name '**Lab4-your last name**' in MATLAB*
- *Save your variables on the workspace at the end of your section or every time that you consider needed, from the file menu → **Save Workspace As** → ---------------*

Downloading the data:

*File > Import data>***L:\PGEG 371 Data Analysis & Geostatistics\Lab Exercises\Lab data\Lab.4\** Students_Height.xls > Next > Finish

Note: Make sure your data is in the Workspace.

Ex. 1 Cutoff & "grade/tonnage" – Students' height data

Problem to solve: Using a cut-off of 1.70 m, find the probability of having the cut-off value

1. **Create your own PDF following the next steps:**

Before doing any calculation in MATLAB, make sure you understand what you are doing, and take in account the following recommendations:

- **Be very careful with the parenthesis.**
- **Use the period "." exactly as is shown below.**

(1) order the heights using the command "sort"

```
>> Height=data
>> Height_order = sort (Height)
```

(2) use the equation of a normal distribution (page 32)

$$\boxed{P(g) = \frac{1}{\sigma\sqrt{2\pi}} e^{\left\{\frac{-(g-\mu)^2}{2\sigma^2}\right\}}}$$

(so, you have to calculate the mean and standard deviation of the sample data, as it is shown in the next two lines)

```
>> Height_std = std(Height,1)
      -------------------
>> Height_mean = mean(Height,1)
      -------------------
```

Then, you can calculate P(g) of your data:

```
>> pdf_Height=(1/ (Height_std *sqrt(2*pi)) ) *exp( -((Height_order-Height_mean).^2 ) / (2*(Height_std^2)))
      -------------------------------------------------------------------------------
```

2. Plot the probability density (Pg):

```
> plot(Height_order,pdf_Height,'r*-')
```

3. Could you make this graph by hand? How? Briefly explain

4. Label the graph that you just displayed (use xlabel and ylabel).

What is in the horizontal axis? (replace ---------- by the horizontal axis name)

```
>> xlabel('-------------')
```

What is in the vertical axis? (replace ---------- by the vertical axis name)

```
>> ylabel('-------------')
```

5. Print the graph, **determine the probability to have a student with a height of 1.70 m** using a RULER, and Show the results on the graph**.**

6. Standardize the cut-off of 1.70 m for the whole population. Show the equation, and write your result.

- Calculate the best estimate for the mean and standard deviation of the complete population.

```
> mu_Height = mean(Height)

                  -------------------

> sigma_Height = std(Height)

                   -------------------
```

X=

7. Assuming that your population has a Normal Distribution, **find the proportion of the population (Φ(x)) that lies above the cutoff**. Use Table 1 that is in:

L:\PGEG 371 Data Analysis & Geostatistics\Lab Exercises\Lab Data\Lab 04\Table_1.xls

Write your result, **Φ(x) =** **=> P=**

Find the standardized density probability (see point/equation c in page 2).

$\phi(x)=$

8. Find the mean of the portion above the cut-off (Bear in mind you are removing a portion from you population, so the mean will be different). What is this new mean? What does this new mean signify? (a) Show the equation, (b) write the result and (c) briefly explain, use one sentence.

Mean_170 =

9. Choose five (5) new different cut-off values based on your data.

Cutoff(1) =

Cutoff(2) =

Cutoff(3) =

Cutoff(4) =

Cutoff(5) =

Make a new variable, which has all cutoff values used.

>> gc = [1.70, Cutoff(1), Cutoff(2), Cutoff(3), Cutoff(4), Cutoff(5)]

10. Find the proportion that lies above the cut-offs, and write them down. (See point a and b in page 2)

$(x)_1 =$	$\Phi(x)_1 =$	=> P_1=
$(x)_2 =$	$\Phi(x)_2 =$	=> P_2=
$(x)_3 =$	$\Phi(x)_3 =$	=> P_3=
$(x)_4 =$	$\Phi(x)_4 =$	=> P_4=
$(x)_5 =$	$\Phi(x)_5 =$	=> P_5=

Make a new variable, which has all proportion above the cutoff values used (substitute the names by the values that you got in previous question).

>> P_gc= [P, P_1, P_2, P_3 , P_4, P_5]

11. Calculate the standardized density probability (see point/equation c in page 2).

$\phi(x)_1 =$

$\phi(x)_2 =$

$\phi(x)_3 =$

$\phi(x)_4 =$

$\phi(x)_5 =$

12. Find the mean for the portion above each cut-off (See point d in page 2)

Mean_cutoff(1) =

Mean_cutoff(2) =

Mean_cutoff(3) =

Mean_cutoff(4) =

Mean_cutoff(5) =

Make a new variable, which has all new means for the corresponding cutoff values used.

```
>> means_gc = [Mean_170, Mean_cutoff(1), Mean_cutoff(2), Mean_cutoff(3)  , Mean_cutoff(4), Mean_cutoff(5)]
```

13. Plot these values as follows:

```
> plotyy(gc,P_gc,gc, means_gc)
```

**** Print this plot**

On the horizontal axis you have the "gc" values, in the right vertical axis you have the new average mean for the group that lie at or above the cutoff and in the left vertical axis you will have proportion that lies above the gc, .

14. Check this plot and analyze what happens at each different cutoff. Write down your comments.

Ex. 2 Cut-off & "grade/tonnage" – Data assigned to you in Lab 3

1. Based on your data set, choose a cut-off. Write down the cut-off for your geosciences data.
   Cut-off =

2. Explain why you select the above cutoff, e.g. you want to find the population taller than 170 cm so I can design the height of the doors.

3. Assuming that your population has a Normal Distribution, find the proportion of the population that lies ABOVE the cut-off, as you did in **Ex. 1.** Show equation and write the result.

x =

Φ(x) = => P=

ϕ(x) =

Mean_cutoff =

4. What is this new mean in terms of your geosciences data? Briefly explain.

Ex. 3. Lognormal, Binomial and Poisson distribution

Use **disttool** command to study these three different distributions, and briefly explain:

1. What is the difference between Normal and Lognormal distribution

PDF

CDF

Ex. 4. What distribution my data follows the best?

Based on what you have learnt in previous labs and in this lab, identify the type of distribution that your data could be best described as:

Student ID & Name:

Post-Lab. Exercise

Due Date: Wednesday 4th March.

Q1. Draw a graph that describes a Normal distribution and explain.

Q2. Draw a graph that describes a Lognormal distribution and explain.

Q3. Group of students come from a population with an estimated average of 175.2cm and an estimated standard deviation of 5.85cm, Calculate the new average for the proportion of the students who are taller than 180cm?