Lab13

October 15, 2020

1 Laboratory 13 Probability Modeling

1.1 Full name:

1.2 R#:

1.3 HEX:

1.4 Title of the notebook

1.5 Date:

1.5.1 Important Terminology:

Population: In statistics, a population is the entire pool from which a statistical sample is drawn. A population may refer to an entire group of people, objects, events, hospital visits, or measurements. **Sample:** In statistics and quantitative research methodology, a sample is a set of individuals or objects collected or selected from a statistical population by a defined procedure. The elements of a sample are known as sample points, sampling units or observations. **Distribution (Data Model):** A data distribution is a function or a listing which shows all the possible values (or intervals) of the data. It also (and this is important) tells you how often each value occurs.

From https://www.investopedia.com/terms https://www.statisticshowto.com/data-distribution/

[0]:
```
### Important Steps:
1. __Get descriptive statistics- mean, variance, std. dev.__
2. __Use plotting position formulas (e.g., weibull, gringorten, cunnane) and␣
↪plot the SAMPLES (data you already have)__
3. __Use different data models (e.g., normal, log-normal, Gumbell) and find the␣
↪one that better FITs your samples- Visual or Numerical__
4. __Use the data model that provides the best fit to infer about the␣
↪POPULATION__
```

2 Estimate the magnitude of the annual peak flow at Spring Ck near Spring, TX.

The file `08068500.pkf` is an actual WATSTORE formatted file for a USGS gage at Spring Creek, Texas. The first few lines of the file look like:

```
Z08068500                       USGS
```

```
H08068500       300637095261000484833 9SW12040102409    409     72.6
N08068500       Spring Ck nr Spring, TX
Y08068500
308068500       19290530  483007              34.30       1879
308068500       19390603    838               13.75
308068500       19400612   3420               21.42
308068500       19401125  42700               33.60
308068500       19420409  14200               27.78
308068500       19430730   8000               25.09
308068500       19440319   5260               23.15
308068500       19450830  31100               32.79
308068500       19460521  12200               27.97
```

The first column are some agency codes that identify the station , the second column after the fourth row is a date in YYYYMMDD format, the third column is a discharge in CFS, the fourth and fifth column are not relevant for this laboratory exercise. The file was downloadef from

https://nwis.waterdata.usgs.gov/tx/nwis/peak?site_no=08068500&agency_cd=USGS&format=hn2

In the original file there are a couple of codes that are manually removed:

- 19290530 483007; the trailing 7 is a code identifying a break in the series (non-sequential)
- 20170828 784009; the trailing 9 identifies the historical peak

The laboratory task is to fit the data models to this data, decide the best model from visual perspective, and report from that data model the magnitudes of peak flow associated with the probebilitiess below (i.e. populate the table)

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

The first step is to read the file, skipping the first part, then build a dataframe:

[6]:
```
# Read the data file
amatrix = [] # null list to store matrix reads
```

```python
rowNumA = 0
matrix1=[]
col0=[]
col1=[]
col2=[]
with open('08068500.pkf','r') as afile:
    lines_after_4 = afile.readlines()[4:]
afile.close() # Disconnect the file
howmanyrows = len(lines_after_4)
for i in range(howmanyrows):
    matrix1.append(lines_after_4[i].strip().split())
for i in range(howmanyrows):
    col0.append(matrix1[i][0])
    col1.append(matrix1[i][1])
    col2.append(matrix1[i][2])
# col2 is date, col3 is peak flow
#now build a dataframem
```

[4]:
```python
import pandas
df = pandas.DataFrame(col0)
df['date']= col1
df['flow']= col2
```

[5]:
```python
df.head()
```

[5]:

	0	date	flow
0	308068500	19290530	48300
1	308068500	19390603	838
2	308068500	19400612	3420
3	308068500	19401125	42700
4	308068500	19420409	14200

Now explore if you can plot the dataframe as a plot of peaks versus date.

[7]:
```python
# Plot here
```

From here on you can proceede using the lecture notebook as a go-by, although you should use functions as much as practical to keep your work concise

[87]:
```python
# Descriptive Statistics
```

[88]:
```python
# Weibull Plotting Position Function
```

[89]:
```python
# Normal Quantile Function
```

[90]:
```python
# Fitting Data to Normal Data Model
```

2.1 Normal Distribution Data Model

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

[91]:
```
# Log-Normal Quantile Function
```

[92]:
```
# Fitting Data to Normal Data Model
```

2.2 Log-Normal Distribution Data Model

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

[93]:
```
# Gumbell EV1 Quantile Function
```

[94]:
```
# Fitting Data to Gumbell EV1 Data Model
```

2.3 Gumbell Double Exponential (EV1) Distribution Data Model

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

[95]:
```
# Gamma (Pearson Type III) Quantile Function
```

[96]:
```
# Fitting Data to Pearson (Gamma) III Data Model
# This is new, in lecture the fit was to log-Pearson, same procedure, but not␣
↪log transformed
```

2.4 Pearson III Distribution Data Model

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)

Exceedence Probability	Flow Value	Remarks
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

[97]:
```
# Fitting Data to Log-Pearson (Log-Gamma) III Data Model
```

2.5 Log-Pearson III Distribution Data Model

Exceedence Probability	Flow Value	Remarks
25%	????	75% chance of greater value
50%	????	50% chance of greater value
75%	????	25% chance of greater value
90%	????	10% chance of greater value
99%	????	1% chance of greater value (in flood statistics, this is the 1 in 100-yr chance event)
99.8%	????	0.002% chance of greater value (in flood statistics, this is the 1 in 500-yr chance event)
99.9%	????	0.001% chance of greater value (in flood statistics, this is the 1 in 1000-yr chance event)

3 Summary of “Best” Data Model based on Graphical Fit