CS 4633 - CS 5623 Simulation Techniques

How to model data using matlab

Instructor Dr. Turgay Korkmaz

This tutorial along with matlab's statistical toolbox manual will be useful for HW 5.

1 Data collection

I got `npd-routes-1.tar.gz` data from `http://ita.ee.lbl.gov/html/contrib/NPD-Routes.html` Basically, they collected data using `traceroute`. After un-tar'ing the data on a linux machine, I used the followings to extract the round-trip-time that I want to model

```
> cat r*/* | awk '{if (NF ==8)  print $1, $3, $5, $7 }' > tall
```

The file `tall` has the following format

```
hopcount time1 time2 time3
```

where `time1, time2, and time3` denote the round-trip-times between the source and the router which is `hopcount` hops away from the source node.

After cleaning some invalid rows manually, I selected the rows with hopcount=15 and saved them into `t15` using

```
> cat tall | awk '{if (($1==15))  print $1, $2, $3, $4 }' > t15
```

Now we have some data to analyze! Start Matlab and do the followings:
We can load our data in matlab as follows

```
>> load t15
```

Assume that we consider only the `time1` data which can be accessed using t15(:,2)
We can compute sample mean, variance, median, min, max for `time1` data

```
>> min(t15(:,2))
>> max(t15(:,2))
>> median(t15(:,2))
>> mean(t15(:,2))
>> var(t15(:,2))
```

To get help about any matlab function use

```
>> help func-name
>> help mean
```

2 Identify the prob dist family

In this part, we need to have a frequency or histogram diagram. For this, matlab has a function called `hist` which (by default) puts data points into 10 equally spaced intervals.

```
hist(t15(:,2));
ylabel('frequency');
xlabel('intervals: values of data points');
```

These three commands will give the histogram in Figure 1. This histogram shows that the distribution of

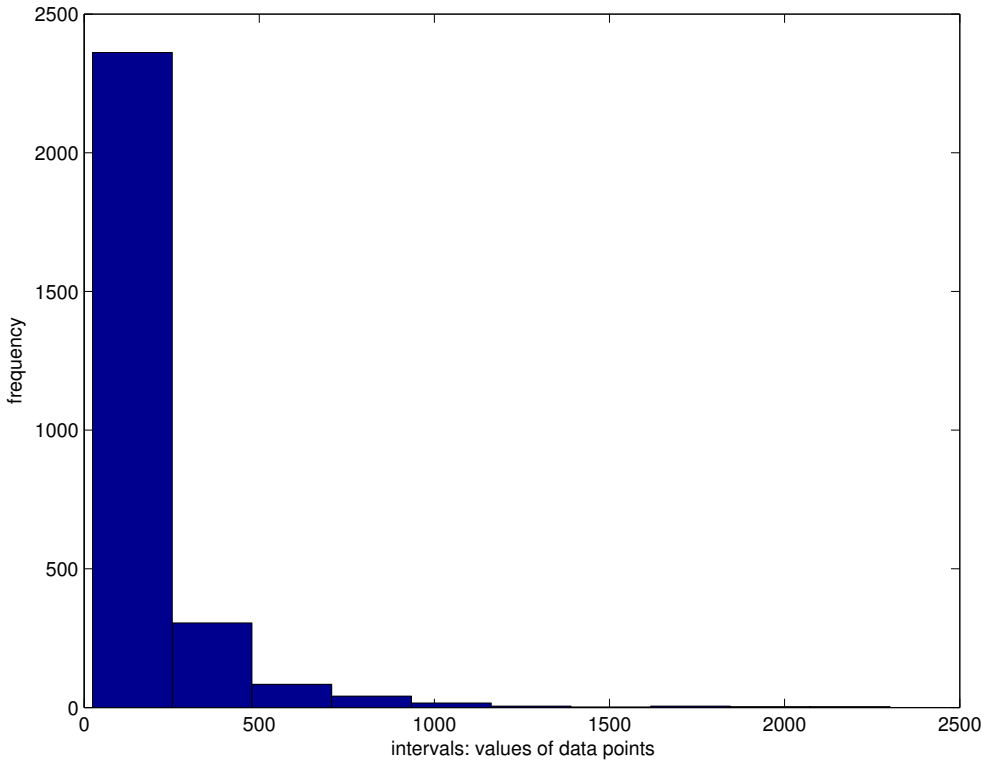

Figure 1: Histogram within 10 intervals (default).

our data looks like an exponential distribution, right! But this will not be a good fit because default `hist` hides so much valuable data.

Remember the recommended number of intervals is $\sqrt{n}$, where n is the number of data points. To determine the number of data points and the number of intervals, we can use the followings in matlab:

```
n=size(t15(:,2));
ninterval=ceil(sqrt(n(1)));
```

Then, we can draw the histogram using

```
hist(t15(:,2),ninterval);
```

These three commands will give the histogram in Figure 2.

Now if you run `disttool` in matlab and consider various distributions and their pdfs, you will see that histogram looks like gamma (Weibull also seems good), as shown in Figure 3. Accordingly, we can select

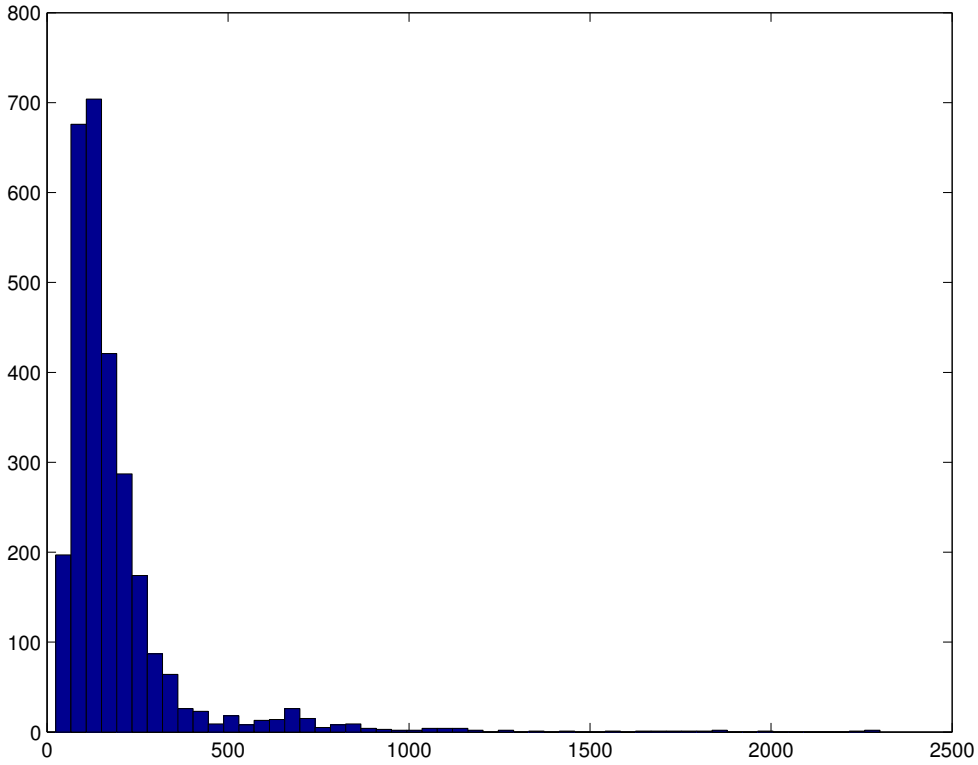

Figure 2: Histogram within $\sqrt{n}$ intervals (recommended).

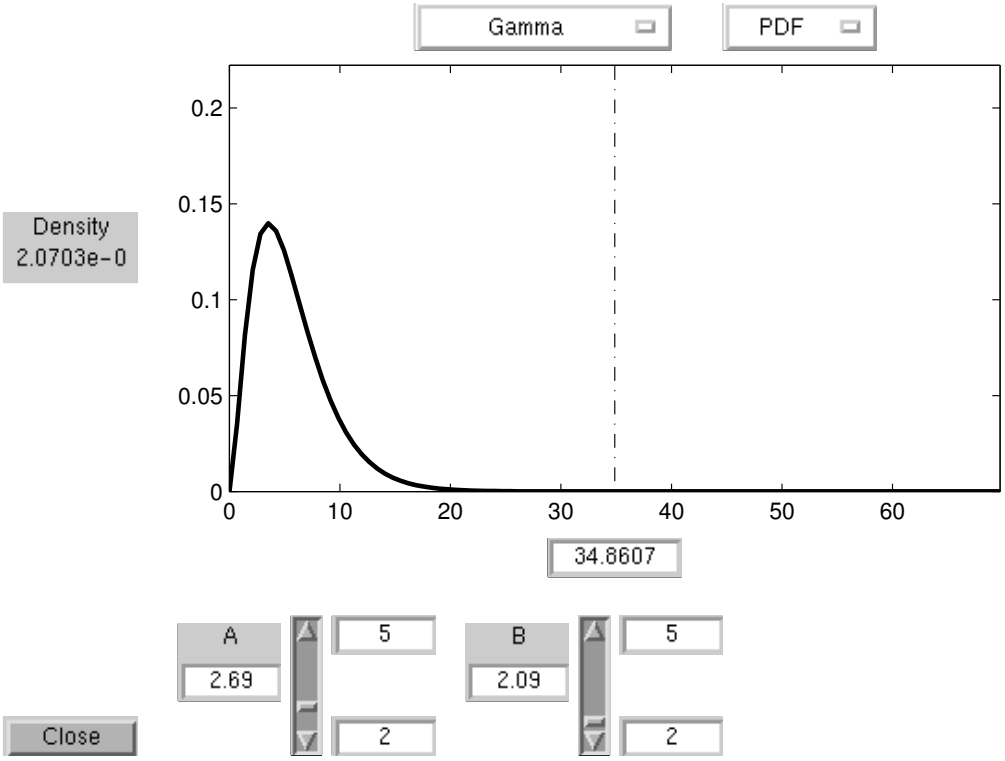

Figure 3: The `disttool` showing the pdf of gamma distribution.

family of gamma dist in this step as the best possible model for our data.

Actually, matlab also provides another nice function to make the first guess accurately. For example, we can also draw the empirical pdf to have better idea in deciding which distribution family to choose. For this, we can use `ksdensity` function as follows.

```
[f,x, u ]= ksdensity(t15(:,2));
plot(x,f,'r')
hold on
[f,x ] = ksdensity(t15(:,2),'width',2*u);
plot(x,f,'b')
[f,x ] = ksdensity(t15(:,2),'width',4*u);
plot(x,f,'g')
[f,x ] = ksdensity(t15(:,2),'width',6*u);
plot(x,f,'y')
```

This function produces the smoothed forms of the empirical distribution, as shown in Figure 4. This figure

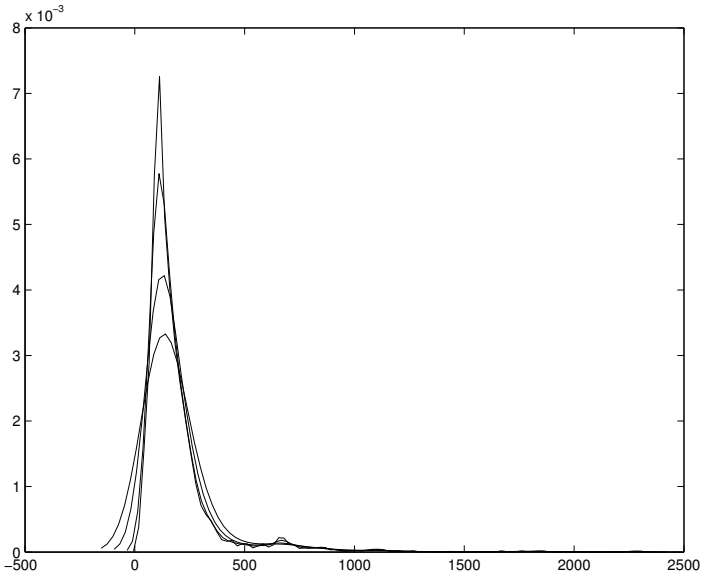

Figure 4: Empirical distribution.

also shows that the empirical pdf of our data looks like gamma distribution, right!.

So, we assume that our data has gamma distribution.

3 Determine parameters for the chosen distribution

Matlab has several functions to estimate parameters of several distributions and the 95% confidence intervals for those estimations.

For our data, if we chose exp, we would use `[muhat, cihat]=expfit(t15(:,2))`

But, since we selected gamma, we will use

```
  [phat, pcihat]=gamfit(t15(:,2))
phat =
    2.1757   87.6674   % means A and B of gamma distribution, respectively.

pcihat =
    2.0682   84.9318
    2.2833   90.4030
```

So, we assume that our data has gamma distribution with
A(θ in our book)=phat(1)=2.1757 and
B(β in our book)=phat(2)=87.6674).

Theoretical pdf of gamma distribution with these parameters can be drawn as follows

```
x=0:0.1:2500;
y = gampdf(x,phat(1), phat(2));
plot(x,y)
```

As a result, we get the pdf of gamma with the given parameters as shown in Figure 5.

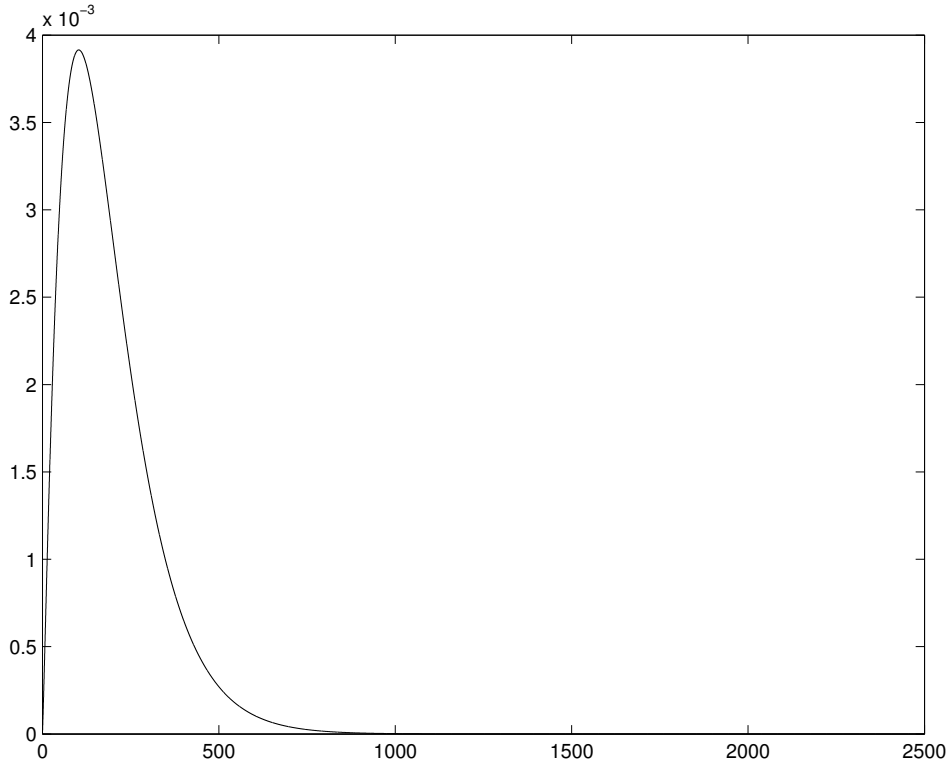

Figure 5: Theoretical pdf of gamma distribution.

4 Evaluate how good the chosen distribution

In matlab, we can do several graphical comparisons between the theoretically assumed distribution and the empirical distribution.

So let's first do graphical comparisons. We can then talk about formal statistical tests.

4.1 Graphical (heuristic) tests

Using matlab, we can draw the theoretical pdf and cdf of gamma dist with estimated parameters A=phat(1) and B=phat(2). We can also draw empirical pdf and cdf. We can then compare them, respectively.

To compare pdfs, we can use

```
[f,x,u ]= ksdensity(t15(:,2));
[f,x ]= ksdensity(t15(:,2),'width',5*u);
plot(x,f,'r')
hold on
x=0:0.1:2000;
y = gampdf(x,phat(1), phat(2));
plot(x,y)
```

As a result, we will get Figure 6.

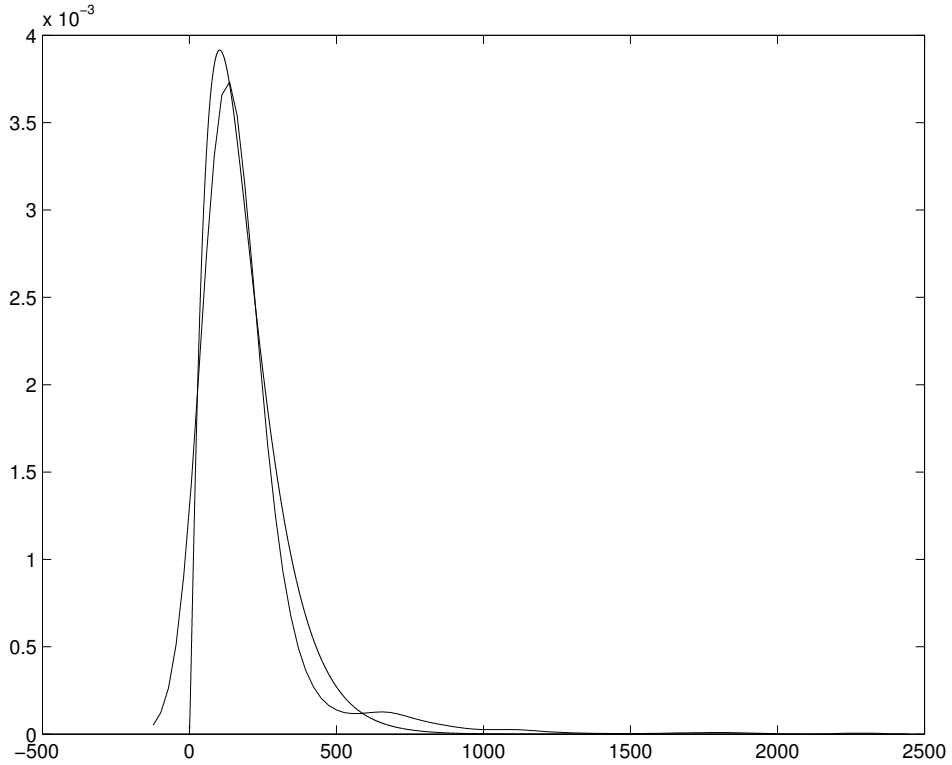

Figure 6: Compare theoretical and empirical pdfs in case of gamma distribution.

To compare cdfs, we can use

```
hold off
[F,x]=ecdf(t15(:,2));
```

```
plot(x,F,'r');
hold on
x=0:0.1:2000;
y=gamcdf(x,phat(1), phat(2));
plot(x,y)
```

As a result, we will get Figure 7.

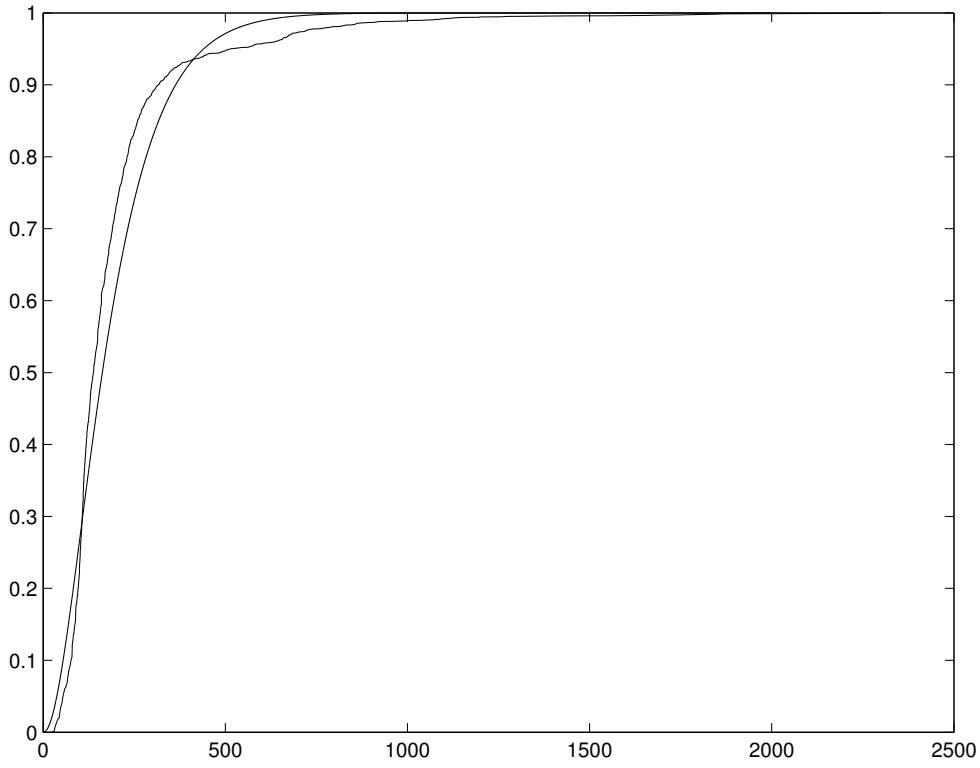

Figure 7: Compare theoretical and empirical cdfs in case of gamma distribution.

How do they look? They seem similar, right! But we cannot trust these similarities. So we need to look at the qqplot, ppplot, scater, to check if the assumed and empirical distributions are really similar

To draw qqplot, you need to use followings.

```
clear
load t15
n=size(t15(:,2));
ninterval=ceil(sqrt(n(1)));
[phat, pcihat]=gamfit(t15(:,2))

x=[];
a=sort(t15(:,2));
prev=-1;
j=0;
for i=1:n(1),
  if prev~=a(i)
    j = j+1;
```

```
    x(j) = a(i);
    prev=a(i);
  end
end
for i=1:j,
  y(i)=gaminv(((i-.5)/j), phat(1), phat(2));
end
plot(x,y,'.');
hold on
```

The above procedure might take too much time. Another simple approach could be the following. Simply, generate some random data from the given theoretical gamma distribution and then use the `qqplot` in matlab to draw the qqplot between the randomly generated data and the empirical data.

```
y = gamrnd(phat(1), phat(2),1,n(1));
qqplot(t15(:,2),y);
legend('Actual procedure','Simple procedure');
```

As a result, we get approximately the same qqplot when we use the actual and simple approach, as shown in Figure 8. As shown in qqplot, the assumed gamma is not a good fit at the tail of the distribution while

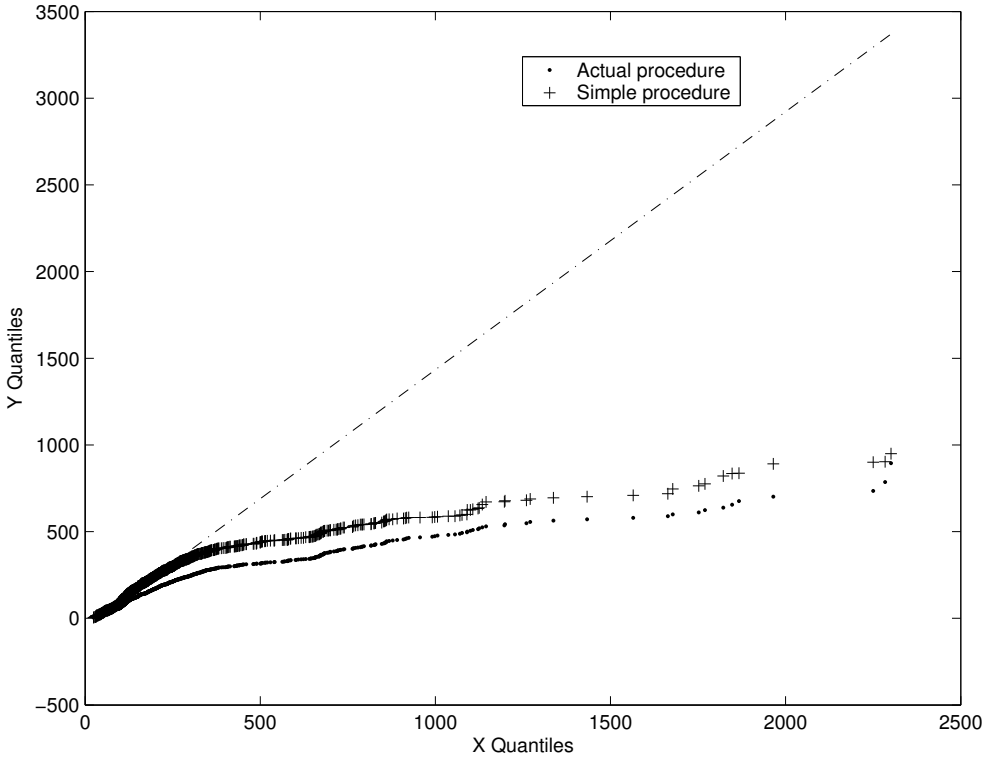

Figure 8: qqplot using actual and a simple procedures.

it is a good fit in the main part of the distribution.

We can also look at the pp-plot. For pp-plot we can use the followings

```
[eF,x]=ecdf(t15(:,2));
```

```
tF =  gamcdf(x, phat(1), phat(2));
plot(eF,tF,'.');
```

As a result, we get the pp-plot in Figure 9.

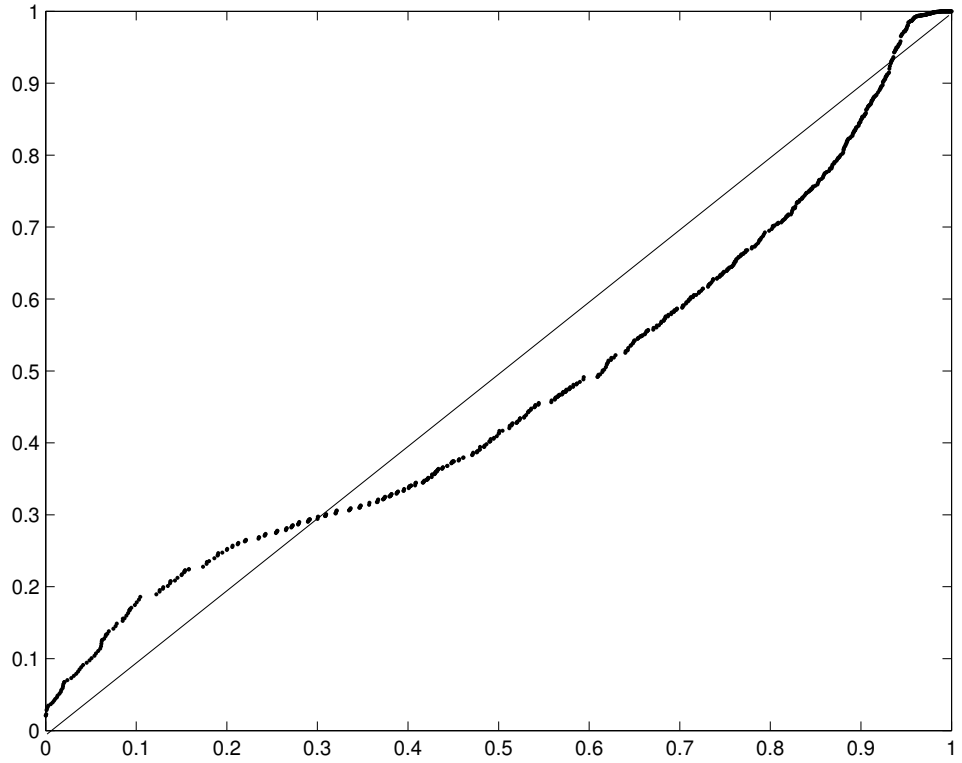

Figure 9: pp-plot.

pp-plot also shows that gamma is a better fit in the middle of dist than the tail. But still the fit is not exact.

At this point, there is no need to use chi-square or KS statistical tests. We need to consider another distribution like Weibull or simply use empirical distribution.

4.2 Goodness-of-fit tests

left as an exercise.

5 Learn by doing it

Download `tall` and then do the followings.

1. For the same data used in this tutorial, consider the Weibull distribution and repeat the above operations.

2. Consider `time1` data when hopcount=10 and repeat the above operations.