Lab Project 4: The Normal Distribution

Course : Introduction to Probability and Statistics, Math 113 Section 3234
Instructor: Abhijit Champanerkar
Date: Nov 14th 2012

Normal Distributions

A large part of statistical analysis is based on the properties of normally distributed random variables. There's a famous LAW that states that data coming from a large number of independent experiments will produce a Normal Distribution and as such, a lot of statistical models assume 'Normality'. We have seen the Empirical Rule for normal distributions before.
The **Empirical Rule** or the **68-95-99.7 Rule** states that for if a frequency distribution of a set of sample data is **normally distributed** then

- Approximately 68% of the data falls within 1 standard deviation of the mean i.e. within $\overline{x} \pm s$.
- Approximately 95% of the data falls within 2 standard deviations of the mean i.e. within $\overline{x} \pm 2s$.
- Approximately 99.7% of the data falls within 3 standard deviations of the mean i.e. within $\overline{x} \pm 3s$.

How *Normal* is a Distribution?

In this project, we will take what we know about normal distributions and compare the theoretical distribution to some real data.
We can find the cumulative area on the left of a z score for a standard normal distribution in **R** using the command:

```
> pnorm(z, mean=0, sd=1)
> pnorm(1, mean=0, sd=1)
[1] 0.8413447
```

Lets load a data set containing information on the air quality in New York.

```
> data(airquality)
> attach(airquality)   ## To allow us to access the named variables by name.
```

The data set contains measurements of

```
> names(airquality)
[1] "Ozone"   "Solar.R" "Wind"    "Temp"    "Month"   "Day"
```

Ozone is the primary ingredient in 'smog'; the more ozone in the air, the worse the air quality. Too much ozone is dangerous to one's health and poses real hazards for the elderly and those with lung ailments. A 'bad air' day is one with high ozone concentrations.
Lets concentrate on the Wind measurements. Since we want to compare these measurements to the Standard Normal Distribution, the first thing we'd like to do is **normalize** the data by converting to . *z-scores*. In R, define a new measurement:

```
> zwind = (Wind - mean(Wind))/sd(Wind)
> hist(zwind, prob=T)
```

Now we can compare the distribution of *zwind* to the Standard Normal Distribution. First, we know that 68% of Normally Distributed data lies between ± 1 standard deviation of the mean. To see if this is true for the Wind data, we need to compute the percentage of data in *zwind* that lies between ± 1. In R:

```
> sum( zwind > -1 & zwind < 1) ## Number of data points in 1 sd of mean
[1] 100
> sum( zwind > -1 & zwind < 1)/length(zwind)  ## Percentage of data in 1 sd
[1] 0.6535948
```

Pretty close. Using the `pnorm` command we find that 86.64 % of Normally Distributed data lies within ± 1.5 standard deviations of the mean. Check this for the Wind data:

```
> sum(zwind > -1.5 & zwind < 1.5)/length(zwind)
[1] 0.869281
```

Again, the Normal Distribution prediction is quite close. Answers the questions below, and enter them on the next page.

Questions

1. As above, find the percentage of Wind data which is less than the following 4 values of z:

   $$-0.75, -1.25, 1.85, 2.85$$

   Compare the percentages in the attached table for the standard normal distribution to the percentages of Wind data found using R.

2. Compute the following Probabilities using (a) the Wind Data and (b) the Normal Distribution. (Note: you will need to convert these to their z–scores).

   (a) Prob(Wind $>$ 10 mph)
   (b) Prob(Wind $>$ 15 mph)
   (c) Prob(Wind $>$ 20 mph)
   (d) Prob(Wind $<$ 5 mph)

3. Take a look at histograms for Ozone, Wind and Temp (do not print). From the histograms, which could best be described by a normal distribution? Give numbers to support your conclusion. Which of the three is 'least' normal (i.e. compare numbers from with Empirical rule) ?

Lab Project 4

Please write your name, fill in the values, tear off and hand to instructor.

Name: _____

Wind data: Question 1

z	-0.75	-1.25	1.85	2.85
Observed Value				
Theoretical Value				

Wind data: Question 2

	Prob(Wind $>$ 10 mph)	Prob(Wind $>$ 15 mph)	Prob(Wind $>$ 20 mph)	Prob(Wind $<$ 5 mph)
z score				
Observed Value				
Theoretical Value				

Question 3