

Stat 324: Lecture 10

Probability plots

Moo K. Chung
mchung@stat.wisc.edu

March 1, 2005

1. *Motivation.* Checking the normality of data is important in setting up a statistical model. Measurements are usually modeled using a normal distribution. We may assume $x_1, \dots, x_n$ are n measurements from $N(\mu, \sigma^2)$. But how do we know the data will follow normality?
2. *Example.* Consider the following example of binge drinking percentage across universities.

```
> library(Devore6)
> data(xmp01.05)
> attach(xmp01.05)
> hist(bingePct, 14)
```

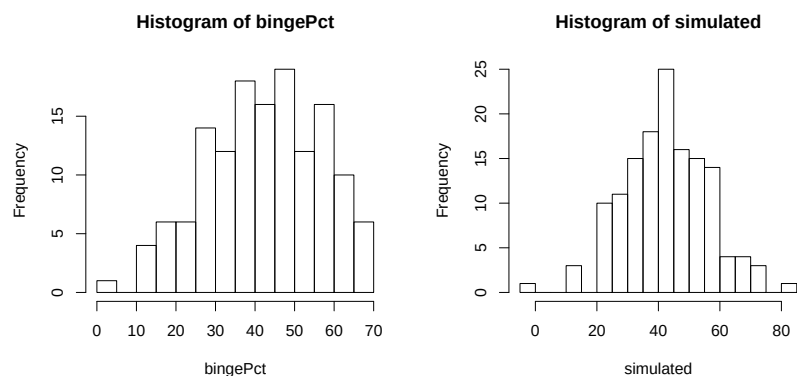


Figure 1: Top: histogram of `bingePct` data. Bottom: histogram of simulated data from $N(42.34, 14.35^2)$.

If it were really from $N(\mu, \sigma^2)$, we can estimate μ and σ from the sample mean and the sample standard deviation:

```
> sd(bingePct)
[1] 14.34699
> mean(bingePct)
[1] 42.33571
> length(bingePct)
[1] 140
```

There are 140 measurement in the data set. Let us randomly generate 140 numbers from $N(42.34, 14.35)$ and plot the histogram.

```
> simulated <- rnorm(140, 42.34, 14.35)
> hist(simulated, 14)
```

Unfortunately eyeball comparison can be deceptive so we need a better method to compare the two histograms. We will compare the percentiles.

3. *Definition.* Given n observations $x_1, \dots, x_n$, we order them from the smallest to the largest and we have $x_{(1)}, \dots, x_{(n)}$. The i -th smallest observation is defined as the $(i - 0.5)/n$ -th *sample quantile* or the $100(i - 0.5)/n$ -th *sample percentile* point. If `bingePct` really follows $N(42.34, 14.35^2)$, then the sample quantiles should be reasonably close to the corresponding quantiles of the normal distribution.

```
> sorted_bingePct <- sort(bingePct)
> sorted_simulated <- sort(simulated)
> plot(sorted_simulated, sorted_bingePct)
```

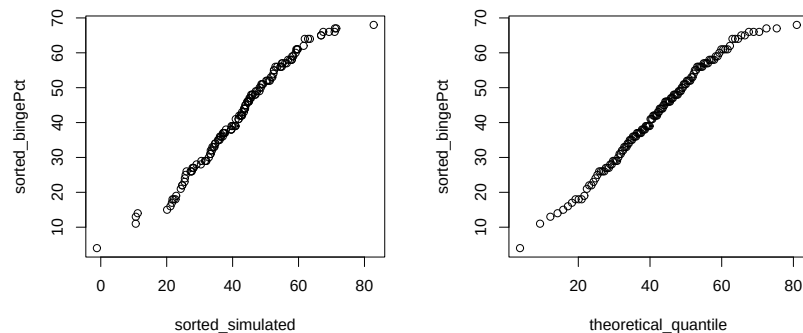


Figure 2: Left: plot of sorted data `bingePct` vs. `sq`-th quantile. This plot is called the *quantile-quantile plot* (QQ-plot) and is used to check the equality of distribution between two samples. This plot shows more or less a straight line so `bingePct` seems to follow normality. Right: normal probability plot of `bingePct`. Quantile points for the x-axis is directly computed from $N(42.34, 14.35^2)$ without simulation.

4. *Normal probability plot.* A better method is to generate the percentile points directly from the theoretical distribution $N(42.34, 14.35^2)$, which can be obtained by

```
> theoretical_quantile <- qnorm((1:140-0.5)/140, 42.34, 14.35)
> plot(theoretical_quantile, sorted_bingePct)
```

R software has a very simple command for generating the normal probability plot:

```
> qqnorm(bingePct)
```

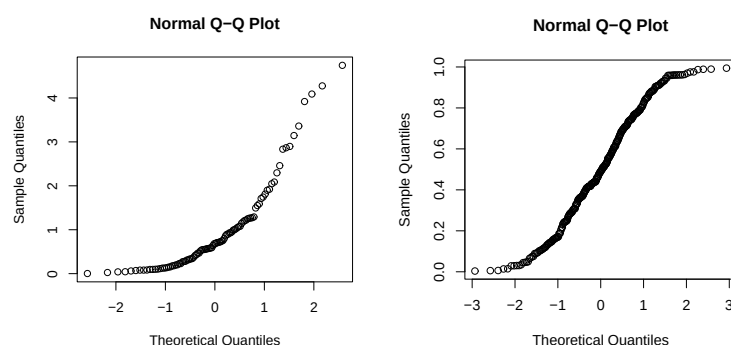


Figure 3: Normal probability plot of data that do not follow normal distributions.

Assigned Problems. Exercise 4.82., 4.88. Homework III due this Thursday. Submit assigned problems from Lectures 07-10.