Chapter 3. Distribution of random variables

Jan 28, 2016

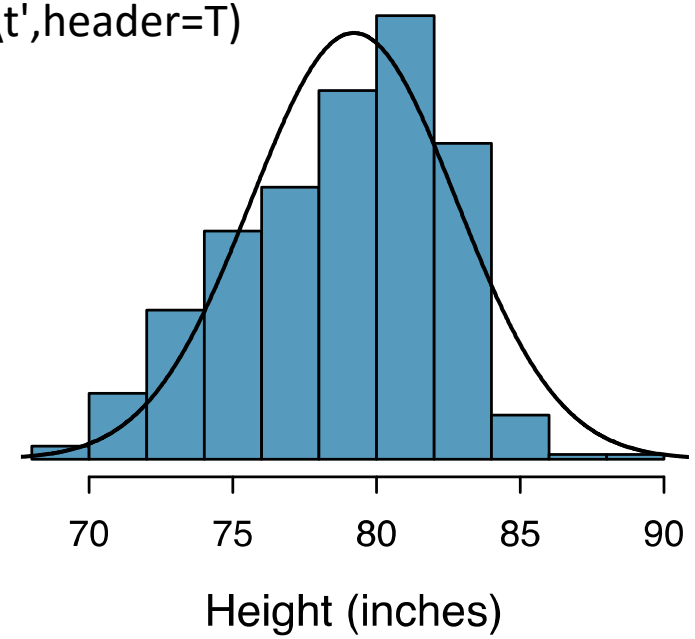
Huamei Dong

1.6. Checking Normality Using Histogram

Many processes can be well approximated by the normal distribution. Using normal Model is also very convenient. But it is always an approximation. So how can we check if the given data is normal distributed? There are two ways of doing that. We learned one way in last lecture---histogram plot. Let's look one more example.

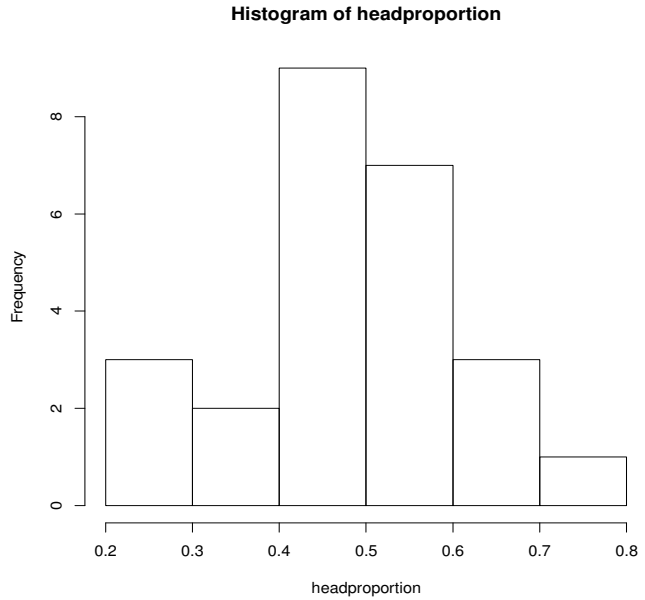
Example 1. Are NBA player heights normally distributed? Consider all 435 NBA players from the 2008-9 season presented in figure 3.12.
Answer: Using R to get the histogram plot.

```
> nba<-read.table("nbaHeights.txt",as.is=T,sep='\t',header=T)
> names(nba)
[1] "last.name"  "first.name" "h.meters"   "h.in"
> hist(nba$h.in)
```



Example 2. To estimate the proportion of “head”, we flip a coin 20 times and count the number of times of getting “head” and then we calculate the proportion of “head”. Repeat this experiment 25 times and plot the histogram.

Answer: The 25 proportions I got are as follows:
0.35, 0.55, 0.3, 0.55, 0.45, 0.45, 0.25, 0.45, 0.55, 0.25, 0.45, 0.6, 0.6, 0.7, 0.55, 0.45, 0.65, 0.5, 0.45, 0.65, 0.45, 0.4, 0.45, 0.55, 0.8
Using R to plot the histogram:
>headproportion<-c(0.35, 0.55, 0.3, 0.55, 0.45, 0.45, 0.25, 0.45, 0.55, 0.25, 0.45, 0.6, 0.6, 0.7, 0.55, 0.45, 0.65, 0.5, 0.45, 0.65, 0.45, 0.4, 0.45, 0.55, 0.8)
>hist(headporportion)



1.7 Check Normality Using Probability Plot

You can also use probability plot to check normality. How to construct a normal probability plot for the head proportion in Example 2?

The idea: We have 25 estimated head proportions. If we arrange the data from the smallest ot the largest, we have

0.25 0.25 0.30 0.35 0.40 0.45 0.45 0.45 0.45 0.45 0.45 0.45 **0.45** 0.50 0.55 0.55 0.55 0.55 0.55 0.60 0.60 0.65 0.65 0.70 0.80

Here for the value 0.45, there are 12 data points below and 12 data points above. So the 50% percentile is 0.45. The percentage for observed value 0.45 is 50%. Generally, we can calculate the percentage like this:

$$\frac{position\ \ in\ \ the\ \ ordered\ \ data}{(\#\ \ total\ \ data\ \ points)\ \ +1}$$

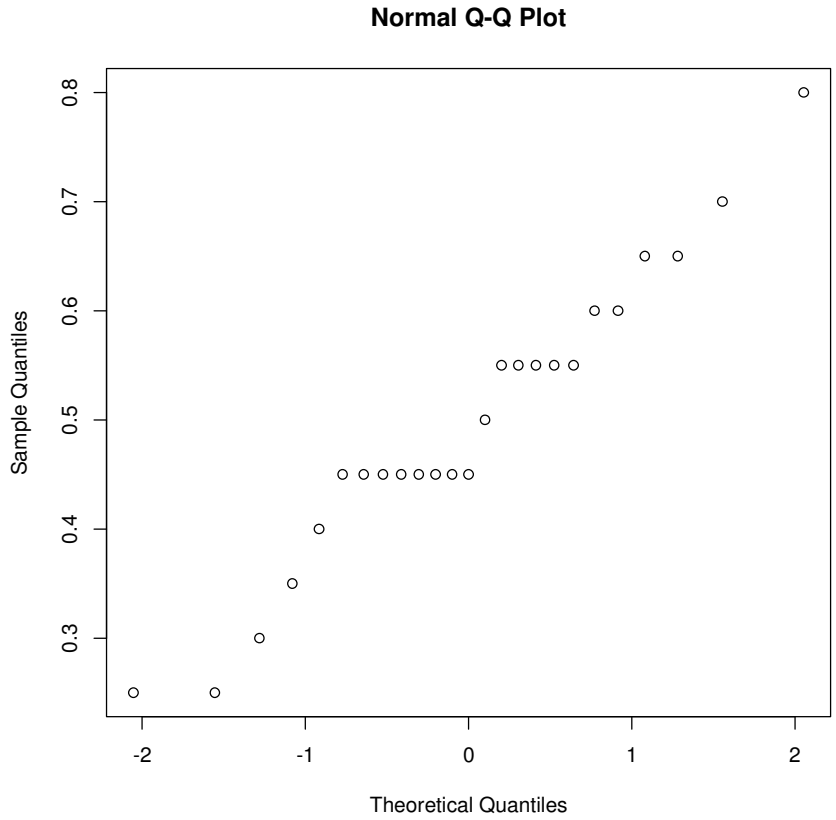
If the sample come from a normally distribution (with mean μ and standard deviation σ), then we should be able to calculate standardized observed value (or Z score) using the percentage and Z table (or using R program). Although we don't know the mean and standard deviation, we know the relation between observed value x and Z score should be linear like this:

$$Z=\frac{x-\mu}{\sigma} \quad or \quad x=\mu+Z\sigma$$

Example 3 Construct the qq plot using the sample head proportion from Example 2.

Answer: Using R

>headproportion<-c(0.35, 0.55, 0.3, 0.55, 0.45, 0.45, 0.25, 0.45, 0.55, 0.25, 0.45, 0.6, 0.6, 0.7, 0.55, 0.45, 0.65, 0.5, 0.45, 0.65, 0.45, 0.4, 0.45, 0.55, 0.8)
>qqnorm(headporportion)



3. Binomial distribution

Example 4. If we flip a coin once, we can get head or tail, 50% of each. If each of five students flips a coin once, what is the probability that we get exactly 2 heads?

Answer: The number of heads is a discrete random variable. This model is binomial model.

Let A, B, C, D, E be the five students. Then we can have two heads from AB, AC, AD, AE, BC,, BD, BE, CD, CE, or DE. There are 10 ways to get two heads.
Probability for just A and B to get the head is $(0.5)^2(0.5)^3$
Similarly probability for AC to get the head is $(0.5)^2(0.5)^3$ and so on.
So finally we have

$$10(0.5)^2(0.5)^3=10(0.5)^5=0.3125$$

Comments: Check if a distribution is binomial , we have to see

(1) The trials are independent
(2) The each trial can be classified as a success or failure
(3) The probability of a success denoted as *p* is the same for each trial

Example 5. Suppose we randomly selected students to participate in a study. We assume 35% students will refuse. What is the probability that 3 out 8 randomly selected students will refuse to participate?

Answer: There are totally $\binom{8}{3}$ many ways of getting 3 refused students.

For each way we have $(0.35)^3(0.65)^5$

So the total probability is $\binom{8}{3}(0.35)^3(0.65)^5$

So the final answer is $\frac{(8)(7)(6)}{(1)(2)(3)}0.35^3 0.65^5 = 56(0.005) = 0.28$

Suppose the probability of a single trial being a success is *p.* Then the probability of Observing exactly *k* successes in *n* independent trials is

$$\binom{n}{k} p^k (1-p)^{n-k}$$

And the mean of the number of successes is $\mu = np$

The standard deviation is $\sigma = \sqrt{np(1-p)}$

3.1. Normal approximation to the binomial distribution

The binomial model is difficult to calculate when sample size n is large. Is there any simpler way to do it?

Normal approximation of the binomial distribution

If a distribution is a binomial distribution with probability of success *p*. When *n* is sufficiently large such that *np* and *n(1-p)* are both at least 10. Then this binomial distribution can be approximated by normal distribution with mean and standard deviation as the following:

$$\mu = np \quad and \quad \sigma = \sqrt{np(1-p)}$$

Example 6. How can we use the normal approximation to estimate the probability of Observing 59 or fewer smokers in a sample of 400 , if the true proportion of smokers is p=0.20?

Answer: Here we have n=400, p=0.20, (1-p)=0.8

So np=80, n(1-p)=320 and we can use normal to approximate it. The mean and Standard deviation for this normal distribution are

$$\mu=np=(400)(0.20)=80, \quad \sigma=\sqrt{np(1-p)}=\sqrt{400(0.20)(0.80)}=8$$

$$P(\text{number of smokers} \leq 59)=P\left(Z \leq \frac{59-80}{8}\right)$$

$$=P(Z \leq -2.63)$$

$$=0.0043$$

If we calculate this probability with binomial model, we are going to do like this

Let x be the number of smokes.

P(x=0) +P(x=1)+P(x=2)+……+P(x=59)

$$= \binom{400}{0}0.2^{0}0.8^{400}+\binom{400}{1}0.2^{1}0.8^{399}+\ldots\ldots\ldots+\binom{400}{59}0.2^{59}0.8^{341}$$

=0.0041

Chapter 3 Homework#4: (due 02/04/16) Page 148 . Exercise 3.43 Page 164. 3.29

Flip a coin 20 times and calculate the head proportion. Repeat this experiment 15 times.

(1) Using histogram to check normality

(2) Using qq plot to check normality