Chapter 3 - The Normal (or Gaussian) Distribution

Read sections 3.1 - 3.2

Basic facts

(3.1)

- The normal distribution gives the distribution for a continuous variable X on the interval (-∞, ∞).
- The notation $X \sim N(\mu,\ \sigma)$ is read as "the variable X is normally distributed with mean μ and standard deviation σ.
- The normal distribution is symmetric, uni-modal, and bell-shaped (see figures on p. 128).
- The normal distribution models many continuous numerical variables in "real life," and hence the term "normal". For example:
  - Heights of human beings are approximately $N(69", 5")$.
  - Lengths of human pregnancies are approximately $N(9$ months, $\frac{1}{2}$ month$)$
  - The density of photons on a mirror after the light is captured by telescope is modeled well by a normal distribution.
  - The size of many different machined parts in industry are modeled by normal distributions.
  - The "erratic motion" of atomic particles is not so erratic, and is well modeled by a normal distribution (*Brownian* motion).

68%, 95%, 99.7% Rule (a special property of the normal distribution)

(3.1.5)

- 68% of the values fall within 1 standard deviation of the mean.
- 95% of the values fall within 2 standard deviations of the mean.
- 99.7% of the values fall within 3 standard deviations of the mean.

Normal probability calculations

We cannot calculate area under the normal distribution like we did for the Witch's Hat and rectangular distributions. We must rely on tabled values (Table B.1 on p. 427) or a software package like R to find probabilities from the normal distribution. Since there are an infinite number of normal distributions, (as many as there are combinations of μ and σ values), then, in order to use a table, we must always transform a normal variable $X \sim N(\mu, \sigma)$ to the **THE STANDARD NORMAL DISTRIBUTION:**

$$Z \sim N(0, 1)$$

The **STANDARDIZATION** transform from X to Z is

$$Z = \frac{X - \mu}{\sigma}.$$

In words, Z is the number of standard deviations that X is from μ!

Probability Calculations about Z:

1. P(Z <1.15) (In R, pnorm(1.15))

2. P(Z > 2.43) (In R, pnorm(2.43,lower.tail=F) OR 1-pnorm(2.43))

3. P(-0.67 < Z < 2.10) (In R, pnorm(2.10) −pnorm(−.67))

"Backwards" Probability Calculations about Z:

1. P(Z < z*) = 0.10 (In R, qnorm(.1))

2. P(Z > z*) = 0.05 (In R, qnorm(.05,lower.tail=F) OR qnorm(1-.05))

3. P(-z* < Z < z*) = 0.95 (In R, qnorm((1-.95)/2)

Probability Calculations about X:

EXAMPLE: Let X be the length of a human pregnancy (in days). The distribution of X is well approximated by the normal distribution. We will use N(266 days, 16 days) to model the distribution of X.

1. P(X < 270) (In R, pnorm(270,mean=266,sd=16)

2. P(X > 300)
   (In R, pnorm(300,mean=266,sd=16,lower.tail=F) OR 1- pnorm(300,mean=266,sd=16))

3. P(240 < X < 280) (In R, pnorm(280,mean=266,sd=16)- pnorm(240,mean=266,sd=16))

"Backwards" Probability Calculations about X:

1. How short are the shortest 2.5% of all human pregnancies?

2. How long are the longest 20% of all human pregnancies?

3. Between what two values do the middle 95% of all human pregnancy lengths lie?

Evaluating and transforming to normality
(3.2)

To use many statistical procedures, it is required that the data is from a normal distribution. What do you do when a data set, $X_1, ..., X_n$, is not from a normal distribution? In many cases, you can "transform the data to normality," yielding transformed data $Y_1, ..., Y_n$ which is normally distributed.

One common approach is to use **BOX-COX POWER TRANSFORMATION**. Let X be the response variable before transformation and let Y be the response variable after transformation. A power transformation consists of raising X to the power λ for each case, as long as $\lambda \neq 0$. When $\lambda = 0$, the natural log transformation is appropriate. Mathematically,

$$Y_i = X_i^{\lambda}, \text{ if } \lambda \neq 0$$
$$Y_i = \ln(X_i), \text{ if } \lambda = 0.$$

The Box-Cox transformation is a slight modification of the power transformation:

$$Y_i = \frac{X_i^{\lambda} - 1}{\lambda}.$$

The reason for such a modification is that

$$\lim_{\lambda \to 0} \frac{X_i^{\lambda} - 1}{\lambda} = \ln(X_i).$$

That is, the natural log transformation is a special case of the power transformation. Typically, λ is chosen from the interval -2 to $+2$. The $\log_{10}$ transformation can be substituted for the natural log transformation because

$$\log_{10}(X) = \log_{10}(e) \times \ln(X) \approx 0.434294 \ln(X) \text{ and}$$
$$\ln(X) = \ln(10) \times \log_{10}(X) \approx 2.3026 \log_{10}(X).$$

That is, the log in one base is just a constant multiple of the log in another base. Multiplying scores by a constant does not change the normality of the distribution.

The process is

1. Check for the need to transform the data $X_1, ..., X_n$. Start by assuming that **the data is normal** and then look for evidence which suggests that the data is NOT normal. In statistics, this initial assumption is called the *null hypothesis*.

2. If the evidence suggests that the data is NOT normal, then a transformation is necessary. Find the appropriate λ to use in the transformation from a Box-Cox plot.

3. Check the effectiveness of your transformation: does the new data look normal?

To check for the need for a transformation in step 1 and effectiveness of the transformation in step 3, use density plots, boxplots, and normal probability plots. Another approach to help judge normality is using the *correlation coefficient* to measure the linearity on the normal probability plot.

- Boxplots and Density Plots
  - If these plots are symmetric, then the evidence fails to suggest that the data are not normal.
  - If these plots are severely skewed, then the evidence suggests that the data are not normal.
- Normal Probability Plot - A plot of the data versus values from a normal distribution.
  - If the points follow a linear pattern, then the evidence fails to suggest that the data are not normal.
  - If the points greatly deviate from a linear pattern, then the evidence suggests that the data are not normal.
- Correlation Coefficient (r) - see Table 7.1 on page 320 in the textbook

Table 7.1 - Values to Which r Can Be Compared to Check for Normality

n	5	10	15	20	25	30	40	50	60	75
Critical r	0.832	0.880	0.911	0.929	0.941	0.949	0.960	0.966	0.971	0.976

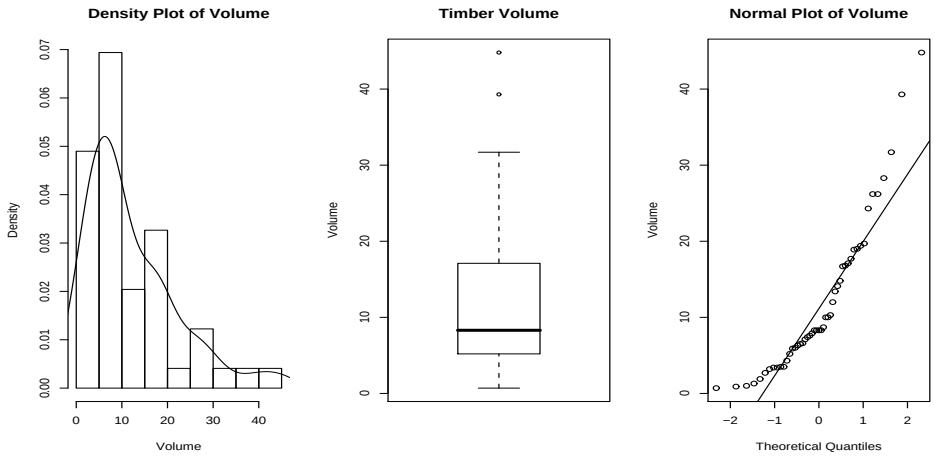
  - The correlation coefficient r quantifies the strength of the linear pattern on the normal probability plot.
    - If r $\geq$ critical r for the given sample size, then the evidence fails to suggest that the data are not normal.
    - If r $<$ critical r for the given sample size, then the evidence DOES suggest that the data are not normal.

EXAMPLE: Timber volume is estimated for 49 different trees. The R-variable name for this data is Volume.

1. Check the data $X_1, ..., X_n$ for normality

- Density plot, boxplot, and normal probability plot

```
> par(mfrow = c(1,3)) # Make three columns in the figure window
> hist(Volume,freq=FALSE,main="Density Plot of Volume",xlab="Volume")
> lines(density(Volume))
> boxplot(Volume,main="Timber Volume",ylab="Volume")
> qqnorm(Volume,main="Normal Plot of Volume",ylab="Volume")
> qqline(Volume)
```



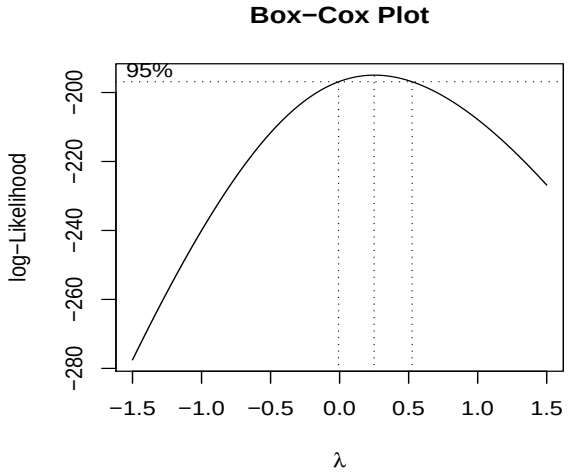
- Calculate r to check for normality:

```
> xy=qqnorm(Volume)
> cor(xy$y,xy$x)
0.9332381
> length(Volume)
[1] 49
```

- Interpretation: From the density plot and boxplot, we see that the data is severely right-skewed. The points in the normal probability plot deviate from a linear pattern. In Table 7.1, our sample size (n=49) is not given. Instead, use the next larger n given, n=50, giving the critical r value of .966. The correlation is r=0.9332 $<$ 0.966. Thus there is evidence to believe the data are not normal.

2. Find an appropriate λ value

```
> library(MASS)
> par(mfrow = c(1,1))
> boxcox(Volume~1,plotit=TRUE,
    lambda=seq(-1.5,1.5,0.01))
> title(main="Box-Cox Plot")
```



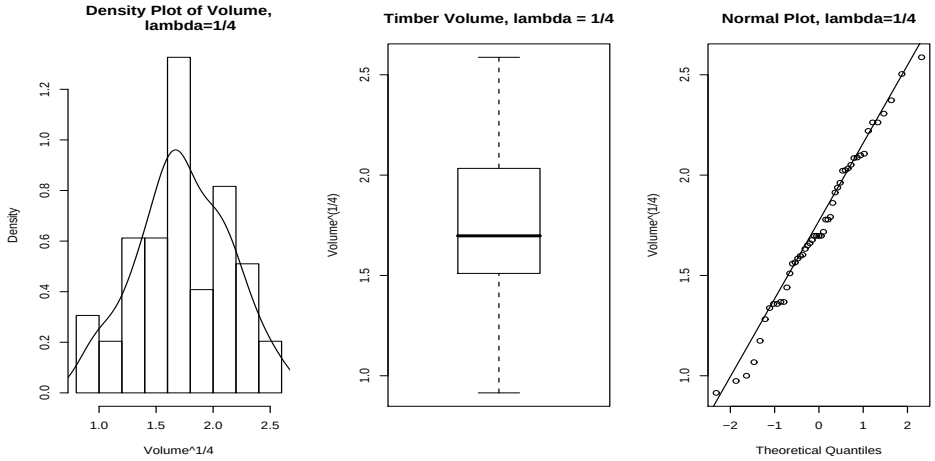
Thus, we'll transform the data $Y = X^{1/4}$.

```
> Volume.new=Volume^(1/4)
```

3. Check the new data $Y_1, ..., Y_n$ for normality

- Density plot, boxplot, and normal probability plot

```
> par(mfrow = c(1,3))
> hist(Volume.new,freq=FALSE,main="Density Plot of Volume, lambda=1/4",xlab="Volume^(1/4)")
> lines(density(Volume.new))
> boxplot(Volume.new,main="Timber Volume, lambda=1/4",ylab="Volume^(1/4)")
> qqnorm(Volume.new,main="Normal Plot, lambda=1/4",ylab="Volume^(1/4)")
> qqline(Volume.new)
```



- Calculate r to check for normality:

```
> xy.new=qqnorm(Volume.new)
> cor(xy.new$y,xy.new$x)
0.994659
```

- Interpretation: From the density plot and boxplot, we see that the transformed data symmetric. The points in the normal probability plot form a linear pattern. The correlation r calculated using the transformed data is $0.9947 > 0.966$. Thus, the evidence does not suggest that the transformed data is not normal (i.e. the transformed data looks normal).

Exercises

Probability calculations on p. 158: 3.1 - 3.13 odd, 3.15a
Evaluating normality on p. 161: 3.17