Chapter 4 Supplement; Sections 6.2 and 6.3

Note made by: Timothy Hanson
Instructor: Peijie Hou

Department of Statistics, University of South Carolina

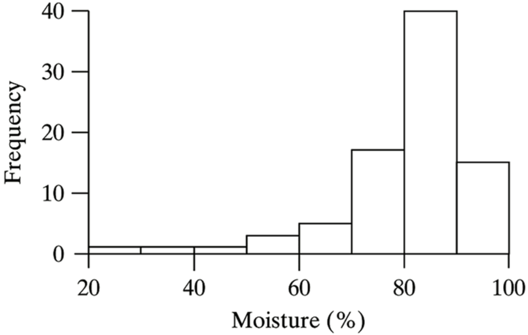
Stat 205: Elementary Statistics for the Biological and Life Sciences

4.4 Checking data are normal

- In many procedures coming up (t tests, confidence intervals, linear regression, & ANOVA) the data are assumed to be normal.
- We'll need to check that assumption.
- Given some data $Y_1, \ldots, Y_n$ we can make a histogram; it should be unimodal and roughly symmetric.
- Your book suggests seeing if data roughly follow the 68/95/99.7 rule. I've never heard of anyone else actually doing this.
- Another option is to make a (modified) boxplot. We expect to see one outlier out of every 150 observations from truly normal data. If we see three or four outliers from a sample of size $n = 50$, the data are not normal.

Example 4.4.2 Moisture content in freshwater fruit

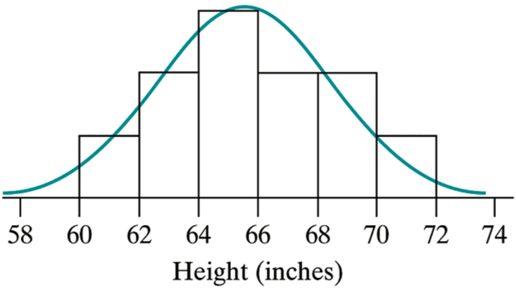
Moisture content was measured in $n = 83$ freshwater fruit. Does the data appear to have come from a normal distribution? Why or why not?



Normal probability plots

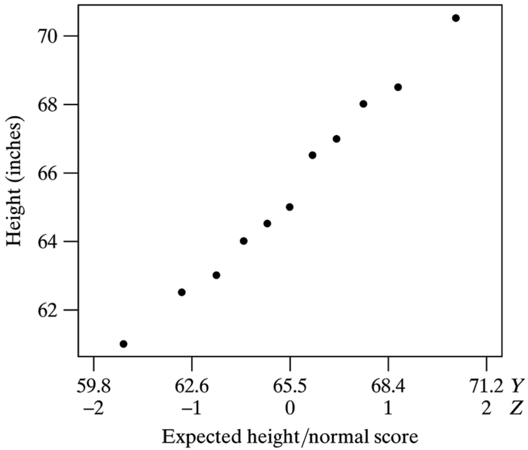
- Another commonly used plot is a normal probability plot or "quantile-quantile" plot.
- $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ is data sorted from smallest to largest.
- The normal probability plot plots the sorted Y_i's against what we'd expect to see from "perfectly" normal data: the percentiles $z_1, \ldots, z_n$ where $\Pr\{Z \leq z_i\} = \frac{i}{n+1}$ for $i = 1, \ldots, n$.
- A computer simply makes a scatterplot of $(z_1, Y_{(1)}), (z_2, Y_{(2)}), \ldots, (z_n, Y_{(n)})$.
- Your book goes into more detail if you're interested.
- These plots will never be perfectly straight due to sampling variability; we're just looking for them to be not totally curved.

Histogram of heights of $n = 11$ women



Histogram with normal density using $\sigma = s = 2.9$ inches and $\mu = \bar{y} = 65.5$ inches. The plot looks okay, but the sample size is pretty small. Let's look at a normal probability plot...

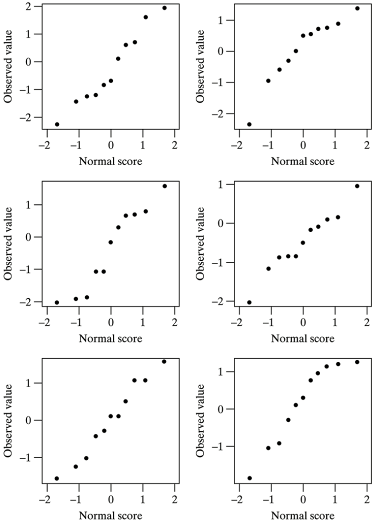
Quantile-Quantile plot of 11 women



The plot is quite straight. The data matches *what we'd expect* from normal data.

Normal probability plots for normal data ($n = 11$)

They're never perfect, but all reasonably straight.



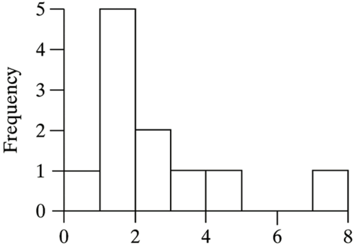
Try it yourself...

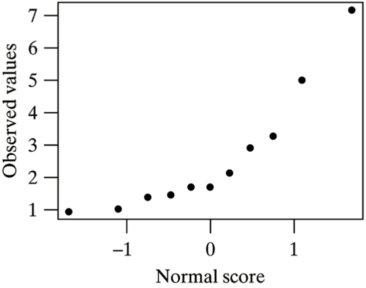
In R type `qqnorm(rnorm(11))` $\boxed{\text{Enter}}$ $\boxed{\uparrow}$ over and over again. Try sample sizes of 50 and 100 too.

In general, if your data set is called, e.g. `heights`, just type `qqnorm(heights)` in R to get the normal probability plot.

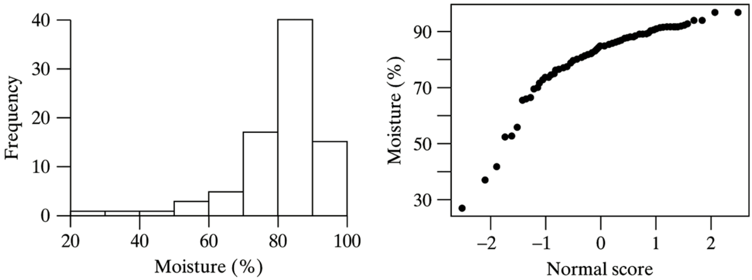
If data *are not normal*, the plot will be non-linear. Let's see some examples.

Data that are skewed right

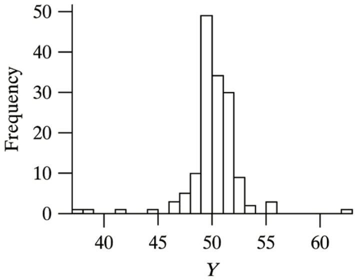


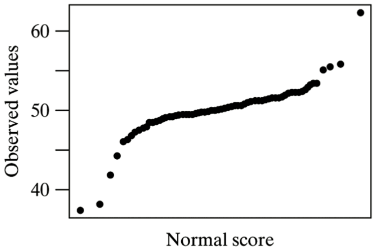


Data that are skewed left

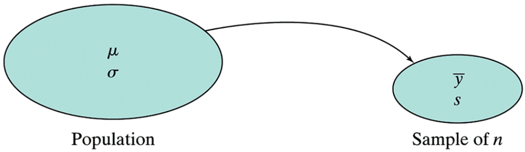


Data with tails fatter than normal





Chapter 6 Confidence interval



Take a random sample of data $Y_1, \ldots, Y_n$ from the population; $\bar{y}$ estimates μ and s estimates σ.

Example 6.1.1 Butterfly wings

$n = 14$ male Monarch butterflies were measured for wing area (Oceano Dunes State Park, California).

Table 6.1.1 Wing areas of male Monarch butterflies

Wing area (cm^2)				
33.9	33.0	30.6	36.6	36.5
34.0	36.1	32.0	28.0	32.0
32.2	32.2	32.3	30.0	

$\bar{y} = 32.81$ cm^2 and $s = 2.48$ cm^2 estimate μ and σ, the mean and standard deviation *of all Monarch butterfly wing areas from Oceano Dunes*.

How good are these estimates? Can we provide a *plausible range* for μ?

6.2 Standard error of $\bar{Y}$

- Recall on p. 151 that $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$.
- We will usually not know σ (if we don't know μ, how can we know σ?)
- Simply plug in s for σ.
- The **standard error of the sample mean** is

$$SE_{\bar{Y}} = \frac{s}{\sqrt{n}}.$$

- For the butterfly wings, $SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{2.48}{\sqrt{14}} = 0.66 \text{ cm}^2$.
- The standard error $SE_{\bar{Y}}$ gives the variability of $\bar{Y}$; the standard deviation s gives the variability *in the data itself.*

Example 6.2.2

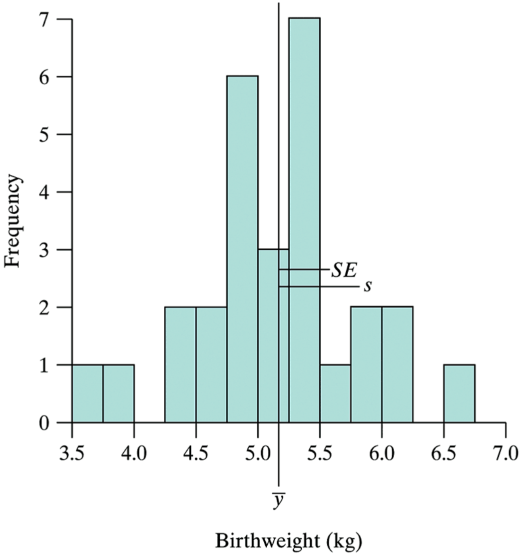
Geneticist weighs $n = 28$ female Rambouillet lambs at birth, all born in April, all single births.

Table 6.2.1 Birthweights of twenty-eight Rambouillet lambs

Birthweight (kg)						
4.3	5.2	6.2	6.7	5.3	4.9	4.7
5.5	5.3	4.0	4.9	5.2	4.9	5.3
5.4	5.5	3.6	5.8	5.6	5.0	5.2
5.8	6.1	4.9	4.5	4.8	5.4	4.7

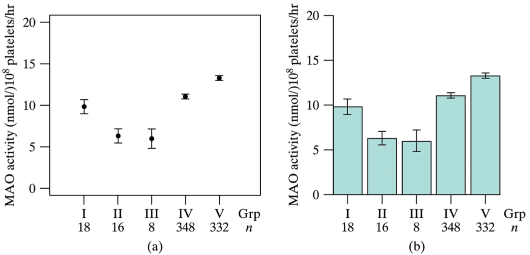
- $\bar{y} = 5.17$ kg estimates μ, the population mean.
- $s = 0.65$ kg estimates the spread *in the sample*.
- $SE_{\bar{Y}} = \frac{s}{\sqrt{n}} = \frac{0.65}{\sqrt{28}} = 0.12$ kg estimates how variable $\bar{y}$ is, i.e. how "close" we can expect $\bar{y}$ to be to μ.

Birthweight of $n = 28$ lambs



Example 6.2.4 MAO data using *SE*'s across groups

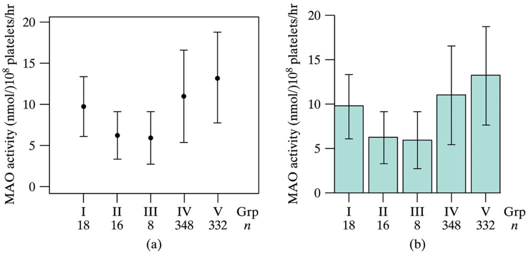
MAO levels vs. schizophrenia diagnosis (I, II, III) and healthy male and female controls (IV and V).



$\bar{y} \pm SE$ using (a) an interval plot, and (b) a bargraph with standard error bars. Gets at how variable the sample means are.

Example 6.2.4 MAO data using s's across groups

MAO levels vs. schizophrenia diagnosis (I, II, III) and healthy male and female controls (IV and V).



$\bar{y} \pm s$ using (a) an interval plot, and (b) a bargraph with standard deviation bars. Gets at how variable the data are.

Example 6.2.4 MAO data table with all information

Table 6.2.2 MAO activity in five groups of people

MAO activity (nmol/10^8 platelets/hr)				
Group	n	Mean	SE	SD
I	18	9.81	0.85	3.62
II	16	6.28	0.72	2.88
III	8	5.97	1.13	3.19
IV	348	11.04	0.30	5.59
V	332	13.29	0.30	5.50

Confidence Interval

- $\bar{y}$ provides an estimate of μ, but it **ignores important information**; namely, how variable the estimator is.
- To avoid this problem (i.e., to account for the uncertainty in the sampling procedure), we therefore pursue the topic of interval estimation (also known as confidence intervals).
- The main difference between a point estimate and an interval estimate is that
  - a **point estimate** is a one-shot guess at the value of the parameter; this ignores the variability in the estimate.
  - an **interval estimate** (i.e., **confidence interval**) is an interval of values. It is formed by taking the point estimate and then adjusting it downwards and upwards to account for the point estimate's variability.

Confidence interval, known σ, formal derivation

Say we know σ (for now) and the data are normal. Then

$$\bar{Y} \sim N(\mu, \sigma_{\bar{Y}}) = N\left(\mu, \frac{\sigma}{\sqrt{n}}\right).$$

We can standardize $\bar{Y}$ to get

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}.$$

We can show $\Pr\{-1.96 \leq Z \leq 1.96\} = 0.95$. Then

Why "confidence"? What if σ is unknown? Non-normal?

- $\bar{Y} \pm 1.96\frac{\sigma}{\sqrt{n}}$ is a 95% probability interval for μ.
- Once we go out and *see* $\bar{Y} = \bar{y}$, e.g. $\bar{y} = 32.8$ cm^2, there is no probability. Either the interval includes μ or not (more in next lecture)
- We don't actually know $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$, but we do know $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$.
- William Sealy Gosset figured out what $\frac{\bar{Y}-\mu}{SE_{\bar{Y}}}$ is distributed as.