Sections 4.3 and 4.4

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

4.3 Areas under normal densities

- Every normal distribution has two parameters μ and σ. These will be given to you in the homework problems.
- A normal random variable with $\mu = 0$ and $\sigma = 1$ is called the **standard normal**, and is denoted Z.
- There is a table of probabilities $\Pr\{Z \leq z\}$ for fixed values of z in Table 3, pp. 616–617.
- Important relationship between $Y \sim N(\mu, \sigma)$ and $Z \sim N(0, 1)$: if Y is normal with mean μ & standard deviation σ,

  $$Z = \frac{Y - \mu}{\sigma}$$

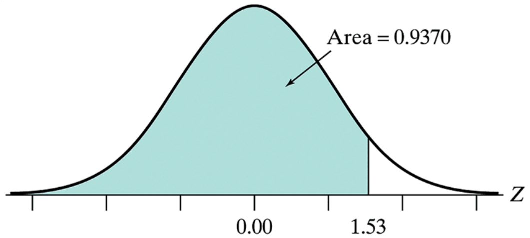
  is standard normal, i.e. normal with mean 0 and standard deviation 1.

Table of standard normal probabilities

- You can get probabilities for any $Y \sim N(\mu, \sigma)$ from Table 3 through "standardization."
- Standardizing Y eventually leads to finding probabilities like $\Pr\{Z \leq z\}$ in Table 3.
- However, computer packages such as R (and online applets) allow computing $\Pr\{Y \leq y\}$ *directly*, so this is the approach I want you to take in homework.
- I'll show you how standardation works anyway, in case you like using tables (and also to explain what the textbook is doing).
- First let's see how to get standard normal $Z \sim N(0, 1)$ probabilities out of the table, and out of R.
- pnorm(y,μ,σ) gives $\Pr\{Y \leq y\}$ for *any* $Y \sim N(\mu, \sigma)$.

$\Pr\{Z \leq 1.53\} = 0.9370$

Along the left side of Table 3 find 1.5, then across the top find the column with 0.03. The intersection of the 1.5 row and the 0.03 column gives the probability 0.9370.



R code:

```
> pnorm(1.53,0,1)
[1] 0.9369916
```

$\Pr\{Z > 1.53\}$

$$\begin{aligned} \Pr\{Z > 1.53\} &= 1 - \Pr\{Z \leq 1.53\} \\ &= 1 - 0.9370 \\ &= 0.0630 \end{aligned}$$

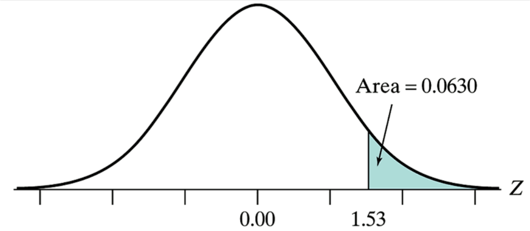


Figure 4.3.3 Area under a standard normal curve above 1.53

R code:

```
> 1-pnorm(1.53,0,1)
[1] 0.06300836
```

$\Pr\{-1.2 \le Z \le 0.8\}$

$$\begin{aligned}\Pr\{-1.2 \le Z \le 0.8\} &= \Pr\{Z \le 0.8\} - \Pr\{Z \le -1.2\} \\ &= 0.7881 - 0.1151 \\ &= 0.6730\end{aligned}$$

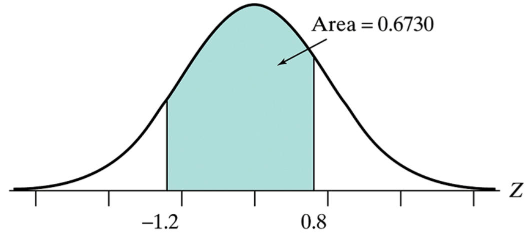


Figure 4.3.4 Area under a standard normal curve between −1.2 and 0.8

```
> pnorm(0.8,0,1)-pnorm(-1.2,0,1)
[1] 0.6730749
```

$\Pr\{Y \leq a\}$ for $Y \sim N(\mu, \sigma)$

$$\begin{aligned}
\Pr\{Y \leq a\} &= \Pr\{Y - \mu \leq a - \mu\} \\
&= \Pr\left\{\frac{Y-\mu}{\sigma} \leq \frac{a-\mu}{\sigma}\right\} \\
&= \Pr\left\{Z \leq \underbrace{\frac{a-\mu}{\sigma}}_{\text{“z-score”}}\right\}
\end{aligned}$$

Now use Table 3.

In R, pnorm(a,μ,σ) does the trick *without standardizing*.

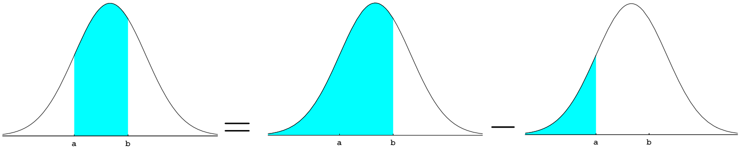
$\Pr\{Y > a\}$ for $Y \sim N(\mu, \sigma)$

$$\begin{aligned} \Pr\{Y > a\} &= 1 - \Pr\{Y \leq a\} \\ &= 1 - \Pr\left\{Z \leq \frac{a - \mu}{\sigma}\right\} \end{aligned}$$

Now use Table 3.

In R, 1-pnorm(a,μ,σ).

Computing $\Pr\{a \le Y \le b\}$ from $\Pr\{Y \le b\}$ & $\Pr\{Y \le a\}$



$$\Pr\{a \le Y \le b\} = \Pr\{Y \le b\} - \Pr\{Y \le a\}$$

$\Pr\{a \leq Y \leq b\}$ for $Y \sim N(\mu, \sigma)$

$$\begin{aligned}\Pr\{a \leq Y \leq b\} &= \Pr\{Y \leq b\} - \Pr\{Y \leq a\} \\ &= \Pr\left\{Z \leq \frac{b-\mu}{\sigma}\right\} - \Pr\left\{Z \leq \frac{a-\mu}{\sigma}\right\}\end{aligned}$$

Now use Table 3.

In R, pnorm(b,μ,σ)-pnorm(a,μ,σ).

“68/95/99.7” rule

For $Y \sim N(\mu, \sigma)$,

- $\Pr\{\mu - \ \sigma \leq Y \leq \mu + \ \sigma\} = 0.68$
- $\Pr\{\mu - 2\sigma \leq Y \leq \mu + 2\sigma\} = 0.95$
- $\Pr\{\mu - 3\sigma \leq Y \leq \mu + 3\sigma\} = 0.997$

This is where the “empirical rule” came from in Chapter 2.

"68/95/99.7" rule for cholesterol in 12–14 year olds

Recall $\mu = 162$ mg/dl and $\sigma = 28$ mg/dl.

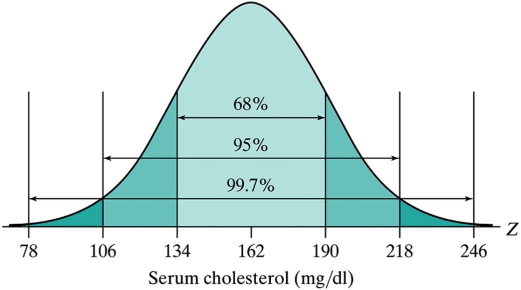
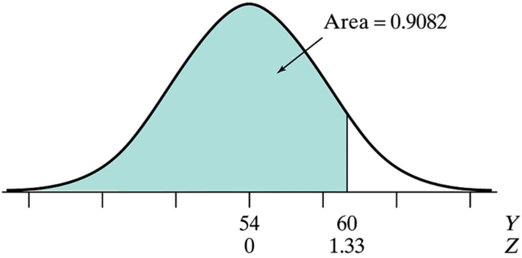


Figure 4.3.6 The 68/95/99.7 rule and the serum cholesterol distribution

Example 4.3.1 Herring lengths

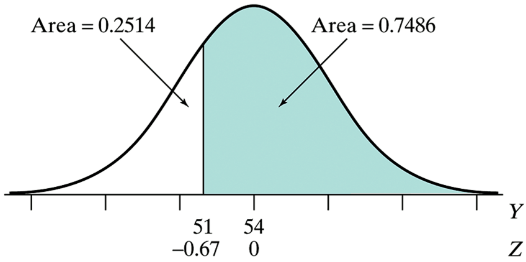
- In a population of herring the lengths of fish are normal with mean $\mu = 54$ mm and $\sigma = 4.5$ mm. Let Y be the length of a randomly selected fish, then $Y \sim N(54, 4.5)$.
- $\Pr\{Y \leq 60\} = \Pr\{Z \leq \frac{60-54}{4.5}\} = \Pr\{Z \leq 1.33\}$ (next slide).
- $\Pr\{Y > 51\} = \Pr\{Z > \frac{51-54}{4.5}\} = \Pr\{Z > -0.67\} = 1 - \Pr\{Z \leq -0.67\}$.
- $\Pr\{51 \leq Y \leq 60\} = \Pr\{-0.67 \leq Z \leq 1.33\}$.

Example 4.3.1(a), $\Pr\{Y \leq 60\}$



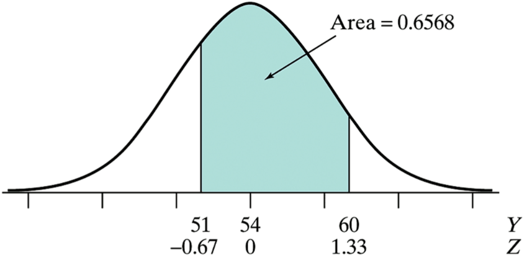
```
> pnorm(60,54,4.5) # using Y ~ N(54,4.5)
[1] 0.9087888
> pnorm(1.33,0,1)  # using Z ~ N(0,1)
[1] 0.9082409
```

Example 4.3.1(b), $\Pr\{Y > 51\}$



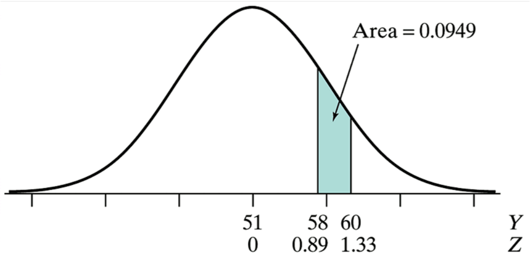
```
> 1-pnorm(51,54,4.5) # direct
[1] 0.7475075
> 1-pnorm(-0.67,0,1) # using z-score
[1] 0.7485711
```

Example 4.3.1(c), $\Pr\{51 \leq Y \leq 60\}$



```
> pnorm(60,54,4.5)-pnorm(51,54,4.5) # direct
[1] 0.6562962
> pnorm(1.33,0,1)-pnorm(-0.67,0,1)  # using z-scores
[1] 0.656812
```

Example 4.3.1(d), $\Pr\{58 \leq Y \leq 60\}$



```
> pnorm(60,54,4.5)-pnorm(58,54,4.5) # direct
[1] 0.09582018
> pnorm(1.33,0,1)-pnorm(0.89,0,1)   # using z-scores
[1] 0.09497381
```

Upper percentile z_α

z_α is defined so that $\Pr\{Z > z_\alpha\} = \alpha$ where $Z \sim N(0,1)$. We'll use this later.

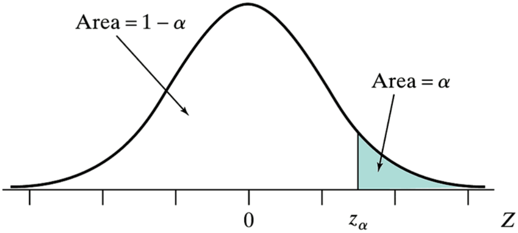


Figure 4.3.12 Area under the normal curve above α

$z_{0.025}$

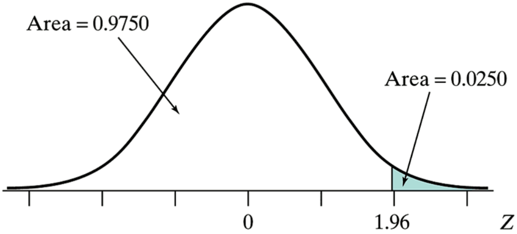


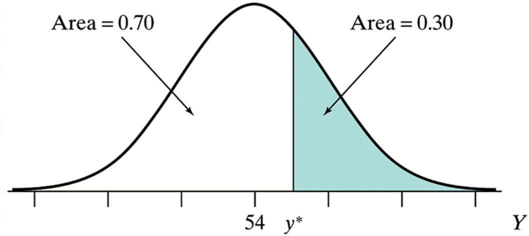
Figure 4.3.11 Area under the normal curve above 1.96

```
> qnorm(0.975,0,1)
[1] 1.959964
```

Percentiles

- For $Y \sim N(\mu, \sigma)$ the number y^* such that $\Pr\{Y \leq y^*\} = p$ is called the $p(100)$**th percentile.**
- These numbers are often used in growth charts, or other biomedical applications where *reference ranges* are needed, i.e. ranges that are "normal."
- You can use Table 3 "in reverse" to get them, but it's easier in R.
- qnorm(p,μ,σ) gives y^*.

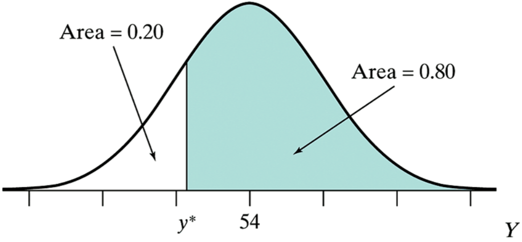
70*th* percentile for Herring size



```
> qnorm(0.7,54,4.5)
[1] 56.3598
```

70% of all Herring are less than $y^* = 56.4$ mm.

20*th* percentile for Herring



```
> qnorm(0.2,54,4.5)
[1] 50.2127
```

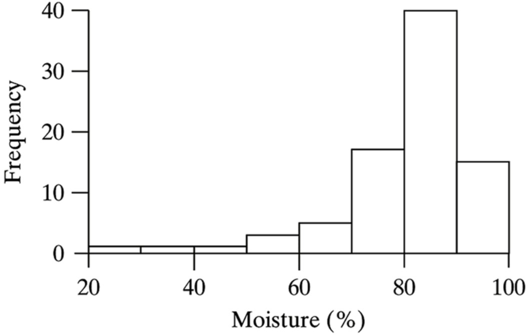
20% of all Herring are *less* than $y^* = 50.2$ mm. 80% of all Herring are *larger* than 50.2 mm.

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 $Y_1, \dots, Y_n$ 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

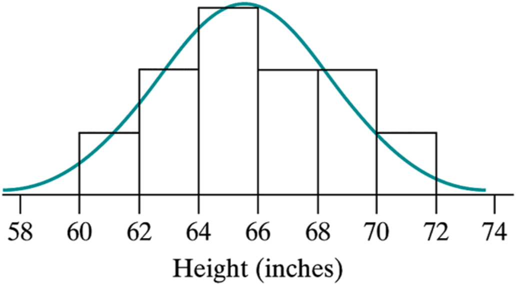
Moisture content was measured in $n = 83$ freshwater fruit. Does the data appear to have come from a normal distribution? Why or why not?



Normal probability plots

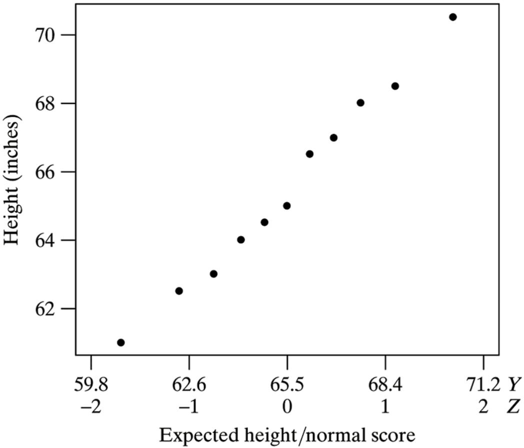
- Another commonly used plot is a normal probability plot or “quantile-quantile” plot.
- $Y_{(1)}, Y_{(2)}, \ldots, Y_{(n)}$ is data sorted from smallest to largest.
- The normal probability plot plots the sorted Y_i's against what we'd expect to see from “perfectly” normal data: the percentiles $z_1, \ldots, z_n$ 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...

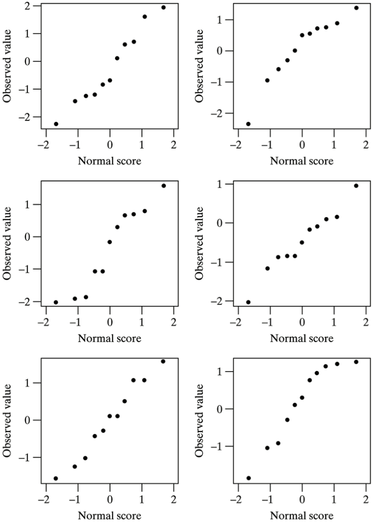
Quantile-Quantile plot of 11 women



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.



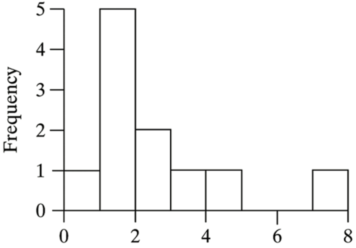
Try it yourself...

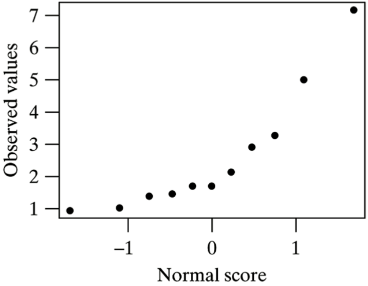
In R type `qqnorm(rnorm(11))` Enter ↑ over and over again. Try sample sizes of 50 and 100 too.

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

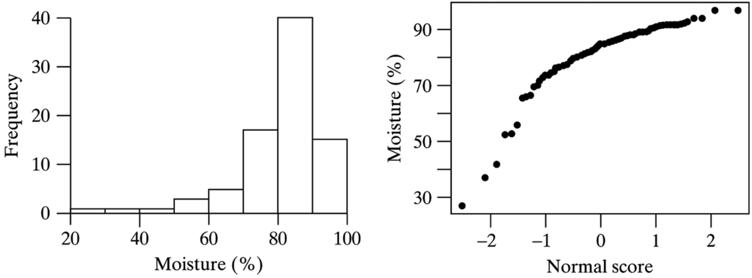
If data *are not normal*, the plot will be non-linear. Let's see some examples.

Data that are skewed right





Data that are skewed left



Data with tails fatter than normal