

February 23, 2009

## Chapter 4: Continuous Distributions

In Chapter 3, the notion of a continuous sample space was introduced. Recall that continuous probability models are useful for modelling responses measured on a continuous scales, such as

- Weights
- Length & Widths
- Volume
- Pressure, etc.

To compute the probability of an event, we cannot add up the probabilities as in the case of a discrete probability example because for continuous distributions the number of sample points is uncountably infinite. Instead, integral calculus is needed to compute probabilities.

### 1 General Concepts

In order to introduce the basic ideas for continuous probability distributions, we introduce an example.

**Example** A study was conducted to differentiate between two different species of voles found in Europe. Several morphometric measurements were obtained from a sample of voles of each species (Airoldi, Flury, and Salvioni 1996). For now we shall look at the variable *skull length* measured in  $\text{mm} \times 100$  of one of the species: *microtus multiplex*. In the table below are  $n = 43$  skull measurements obtained from a random sample of *microtus multiplex* voles arranged from smallest to largest.

2145	2237	2250	2270	2300	2300	2305	2305	2305	2330
2330	2340	2345	2345	2345	2350	2350	2352	2355	2355
2370	2370	2370	2385	2385	2388	2390	2396	2410	2435
2445	2452	2457	2465	2470	2470	2475	2500	2500	2525
2535	2590	2600							

**Table 1.** Skull lengths of  $n = 43$  *microtus multiplex* voles.

Below is a histogram (created using SAS) of the data showing the shape of the distribution.

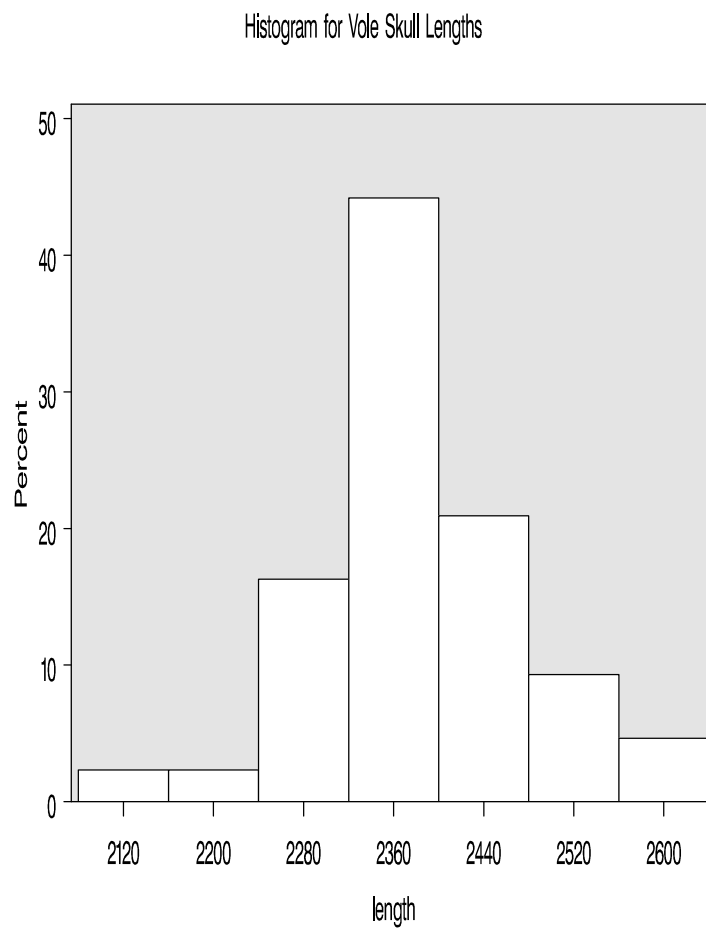


Figure 1: Histogram of the skull lengths of a sample of  $n = 43$  *microtus multiplex* voles.

This histogram shows a nice, symmetric, unimodal “bell-shape.” We can define a random variable  $X$  to be the skull length of a randomly chosen vole.  $X$  is an example of a *continuous random variable* because length is a continuous variable. How do we compute probabilities for this random variable? For instance, suppose we want to know how likely it is that a vole of this species has a skull length of 2600 or greater? From the histogram, it does not appear to be very likely since very few of the voles have skull lengths in this range. What we want to determine then is

$$P(X \geq 2600).$$

Note that we are wanting to determine the probability that the continuous random variable  $X$  assumes values in an interval of real numbers. Another way to regard the probability  $P(X \geq 2600)$  is as the proportion of voles in the population that have skull lengths exceeding 2600. One way to estimate this probability is to look at the proportion of skull lengths in our sample that are 2600 or greater – there is only one skull length in this range:  $1/43 \approx 0.0233$ . However, this method of determining probabilities is not very reliable. For instance, suppose we had not picked the vole with a skull length of 2600 in our sample. Then there would not be any voles in the sample with a skull length of 2600 or greater and therefore the proportion of voles with a skull length of 2600 or greater would be zero. Clearly, zero is not a reasonable estimate of this probability.

Instead of estimating probabilities using a proportion, we can take another look at the histogram and notice that it has a nice “bell-shape”. We can use this shape to propose a probability model for the skull length distribution. Below in Figure 2, a continuous probability *density* function is overlaid with the histogram. In particular, the density curve is the *normal* density function which we will define shortly. The probability density curve can then be used as a model to computing probabilities. High probabilities are associated with high values of the density function. Low probabilities are associated with low values of the probability density function. Probabilities using the density function are determined by computing the *area under the density function*. Since the total probability must be one, the total area under the density curve must also be one. Also, because probabilities cannot be negative, probability density functions can only take nonnegative values.

**Definition: Probability Density Function (pdf)** of a continuous random variable  $X$  is a function that satisfies the following three properties:

1.  $f(x) \geq 0$  for all real numbers  $x$ .
2. For any two real numbers  $a < b$ ,  $P(a < X < b)$  is equal to the area under the graph of  $f(x)$  between these two points. This is illustrated in Figure 4. For those who have had calculus, we know that computing areas under curves requires that we integrate the pdf  $f(x)$ :

$$P(a < X < b) = \int_a^b f(x)dx.$$

For many well-known pdf’s, probabilities can be computed using statistical software packages or tables.

3. The total area under the graph of  $f(x)$  must be equal to one.

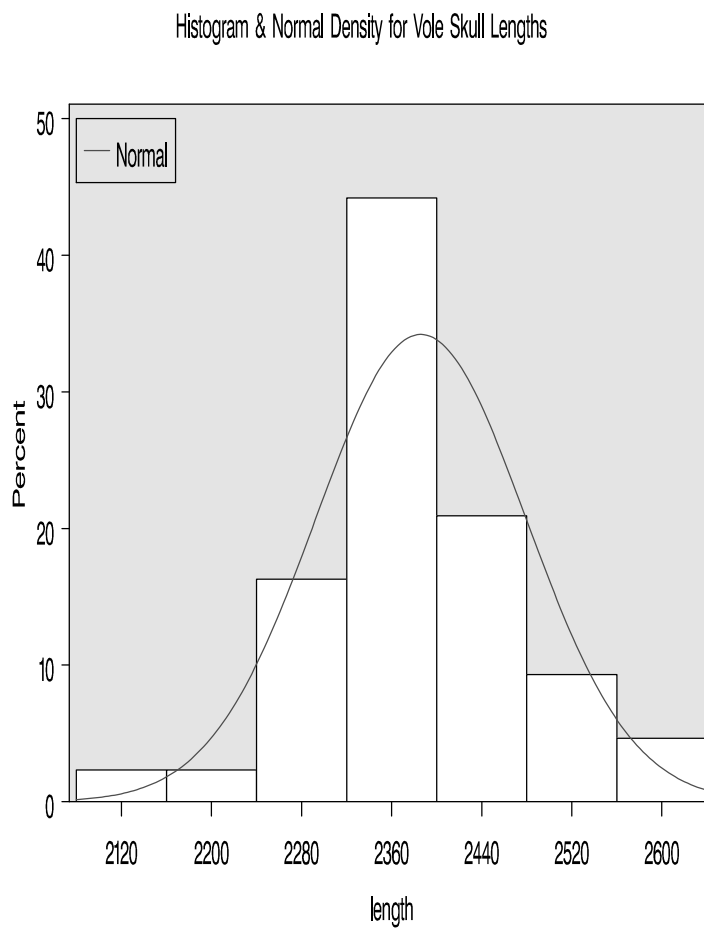


Figure 2: Histogram and a normal probability density overlaid of the skull lengths of a sample of  $n = 43$  *microtus multiplex* voles.

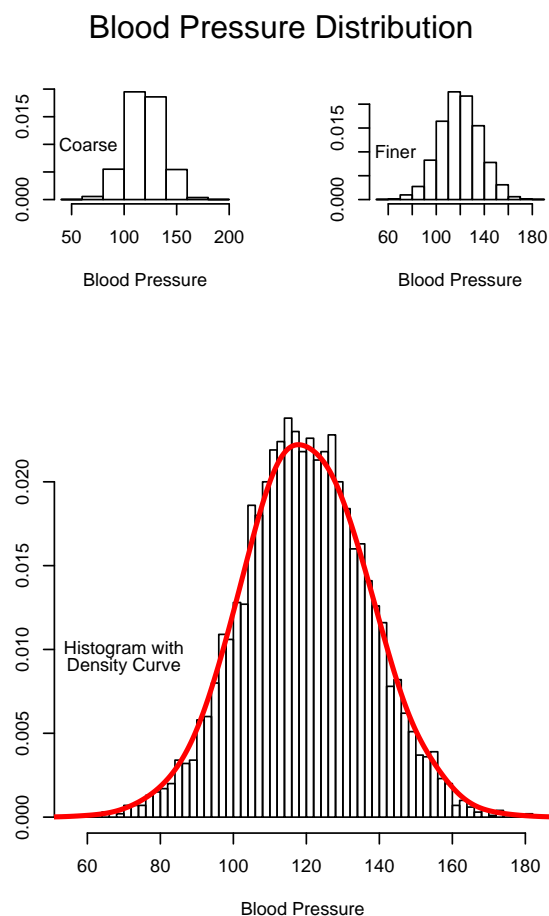


Figure 3: Histograms of simulated data on blood pressures. For a very large sample, we can generate histograms with more and more measurement classes that will approximate the true density curve.

Figure 3 shows histograms for a large simulated data set of blood pressures. We can form a coarse histogram with only a few measurement classes. However, since the sample size is large, we can form histograms with many measurement classes which will reveal the shape of the underlying density function.

One of the important distinctions between continuous and discrete random variables is that for a continuous random variable  $X$ , we have  $P(X = a) = 0$  for any constant  $a$ . The reason for this is that the area under the density at a single point is zero. However, discrete random variables, like the binomial, can associate positive probabilities with an event like  $\{X = a\}$ . For the vole example, we have  $P(X = 2500) = 0$ . If we could measure the length with infinite precision, then no two voles would have exactly the same skull length. Because we are assuming a theoretically infinite population, the proportion of voles with exactly the same skull length would be zero. In practice, measuring instruments can only measure to a certain degree of precision (maybe to the closest millimeter in the skull example). Therefore, all data is measured on a discrete scale even if the theoretical model is continuous.

**Expectation and Variance.** Given a continuous (or discrete) random variable, we can compute its average value, known as the *expectation*. We can also compute the *variance* of the random variable, which is a measure of spread.

**Definition:** The **Expected Value** of a random variable  $X$ , denoted by  $E[X]$  or  $\mu$ , is the average value the random variable can assume. In the discrete case, the expected value was simply a weighted average. In the continuous case, calculus is needed to give a formal definition:

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx,$$

where  $f(x)$  is the density function for  $X$ . For those not familiar with calculus, one can regard the integral sign  $\int$  as a summation over infinitesimally small intervals and thus, the expectation of a continuous random variable can be thought of as a weighted average of the values the random variable can assume, weighted by the probability density function  $f(x)$ . If one plots the probability density on a seesaw, then the seesaw will balance at exactly the expected value. The expected value or mean of a random variable is the center of gravity of the distribution.

**Definition.** The **Variance** of a random variable  $X$ , denoted by  $\text{Var}(X)$  or  $\sigma^2$ , is the expected value (or average) value of  $(X - \mu)^2$ :

$$\sigma^2 = \text{Var}(X) = E[(X - \mu)^2].$$

The calculus definition of variance is given by

$$\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 f(x)dx.$$

From a practical point of view, the variance  $\sigma^2$  of a random variable is almost never known exactly in practice since it has to be estimated from a sample from the full population. However, it can be shown that the sample variance  $S^2$  is a consistent and unbiased estimator of the population variance. This means that as the sample size gets larger and larger, the sample variance  $S^2$  gets closer and closer in value to the population variance with high probability and the sample variance will not systematically over, nor under-estimate, the true population variance.

The (positive) square root of the variance of  $X$  is defined as the **Standard Deviation** of  $X$ , denoted by  $\sigma$ .

One can show with a little algebra the following shortcut formula for computing variances:

$$\text{Var}(X) = E[X^2] - \mu^2.$$

An interesting implication of this formula is that, since the variance cannot be negative, it is always the case that

$$E[X^2] \geq \mu^2.$$

We now turn to the most important continuous probability distribution.

## 2 The Normal Distribution

The *normal distribution*, also known as the *Gaussian* distribution in honor of Karl Friedrich Gauss (1777-1855) is a continuous probability distribution defined by its pdf:

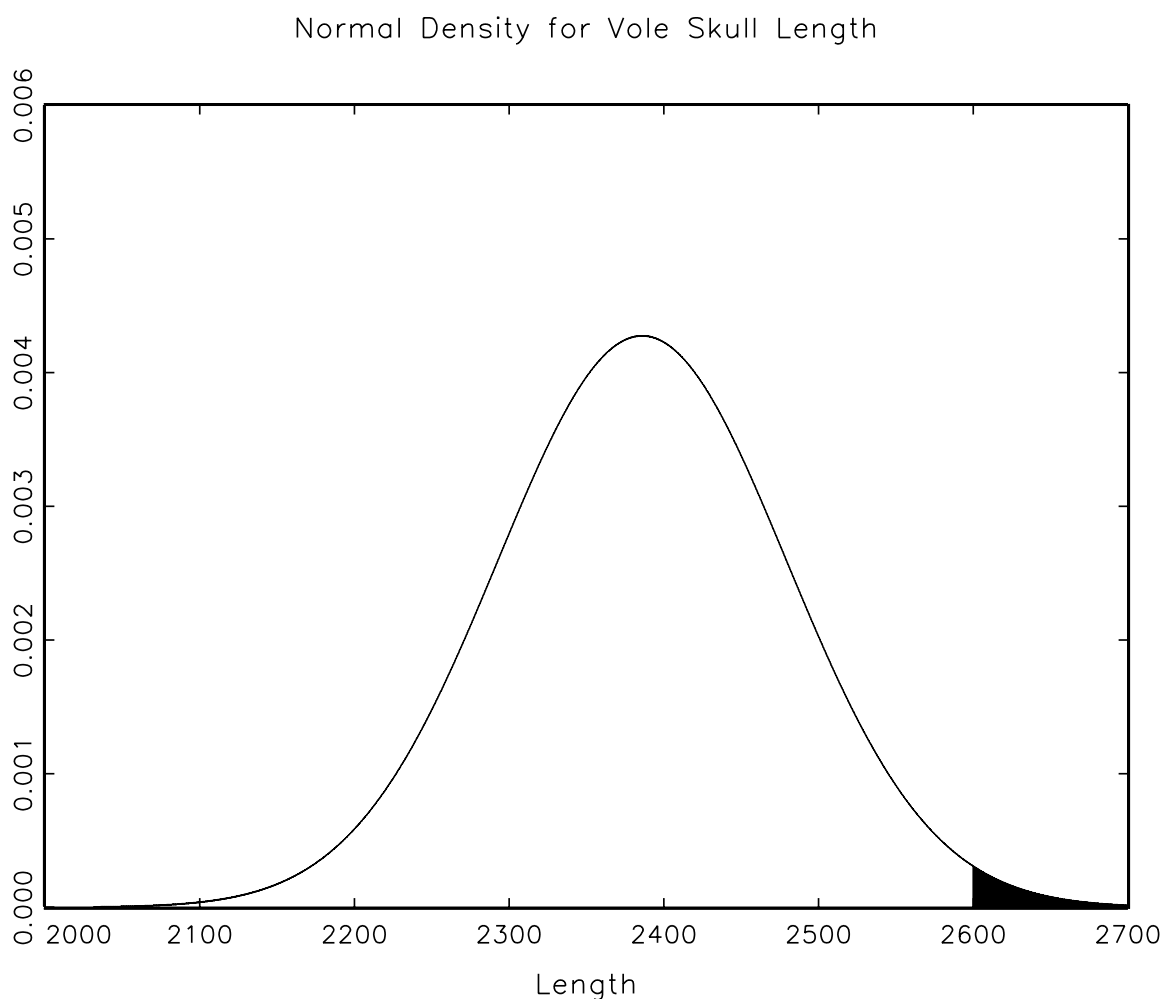


Figure 4:  $P(X > 2600)$  for the vole skull lengths.

**Definition.** We say that a random variable  $X$  has a **normal distribution** with mean  $\mu$  and variance  $\sigma^2$  if its pdf  $f(x)$  is

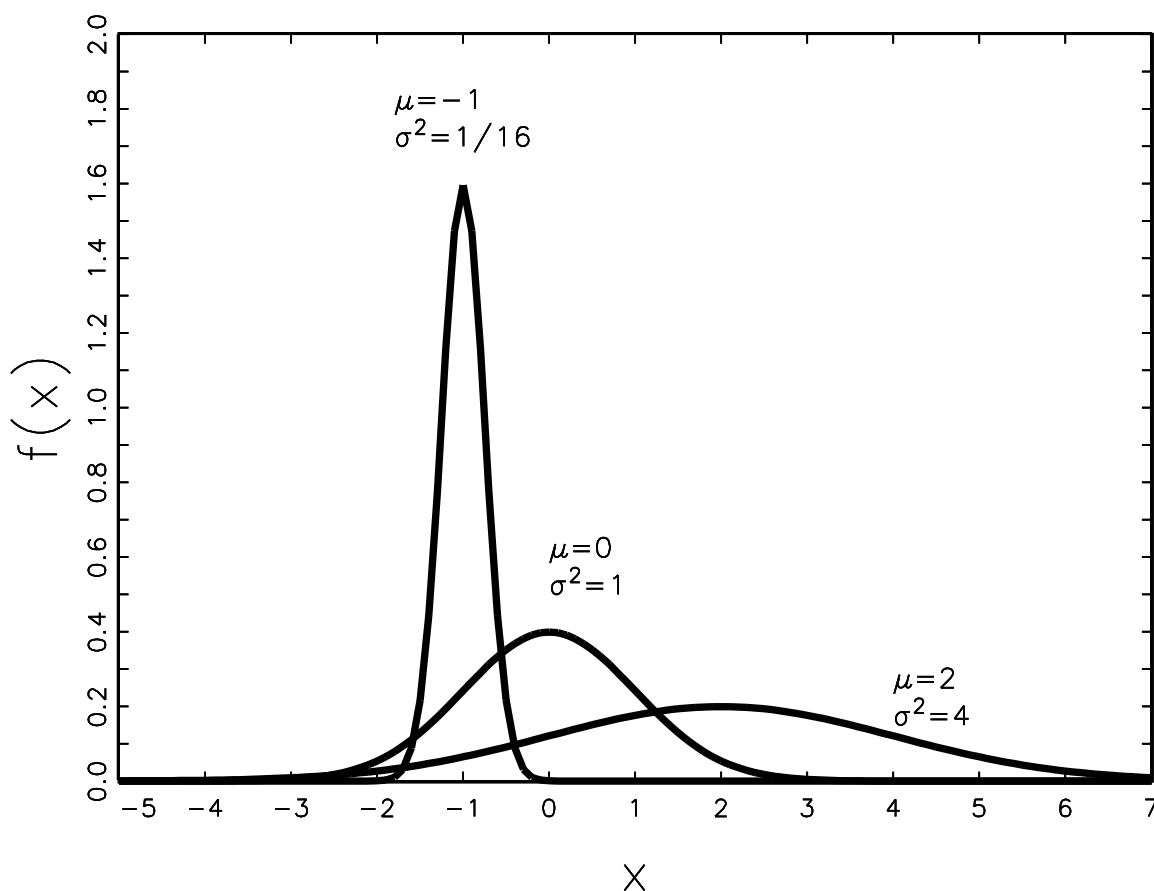
$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(x-\mu)^2/(2\sigma^2)}, \quad -\infty < x < \infty.$$

The normal distribution is denoted by  $N(\mu, \sigma^2)$ . The constant  $\frac{1}{\sqrt{2\pi}\sigma}$  is a normalization constant to make the total area under the curve equal to one. The graph of the normal pdf is a symmetric bell-shaped curve centered at  $\mu$ . The value of  $\mu$  can be any real number and the value of  $\sigma$  can be any positive real number. Figure 2 shows a normal density superimposed over the histogram of the skull length data.

Figure 4 shows the normal density for the vole data using  $\hat{\mu} = 2385.98$  and  $\hat{\sigma} = 93.32$  estimated from the data. The shaded region corresponds to the probability that a skull length will exceed 2600.

**The Role of  $\mu$  and  $\sigma$ .** Changing the value of  $\mu$  in the normal density amounts to shifting the normal density, i.e. changing the location. Making  $\sigma$  bigger (more variability) makes the normal density mound more spread out and not as tall, whereas, making  $\sigma$  smaller causes the normal density mound to become a very steep and tall looking mountain. Figure 5 illustrates the variety of different normal pdf's for varying

## 3 Normal pdf's

Figure 5: Normal pdf's for different values of  $\mu$  and  $\sigma$ 

values of  $\mu$  and  $\sigma$ .

For another illustration, data on the heights of male and female painted turtles was collected. Based on sample statistics, the mean and standard deviation for the males are  $\hat{\mu}_m = 40.7$  and  $\hat{\sigma}_m = 3.36$  mm respectively, while the mean and standard deviation for the female turtles are  $\hat{\mu}_f = 52.0$  and  $\hat{\sigma}_f = 8.16$  mm respectively. Assume also that the height distributions in each population are normal. Then since the female mean is bigger than the male mean, the female density is shifted to the right of the male density curve. Also, since there is more variability in the female heights than in the male heights (i.e.  $\sigma_f > \sigma_m$ ), the female density is more spread out than the male density curve. These observations are apparent in Figure 6 below.

**The Standard Normal Distribution.** A normal random variable with  $\mu = 0$  and  $\sigma = 1$  (i.e.  $N(0, 1)$ ) is said to have a *standard normal distribution* and it will be denoted by  $Z$ .

The density function for the standard normal is

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \quad -\infty < z < \infty.$$

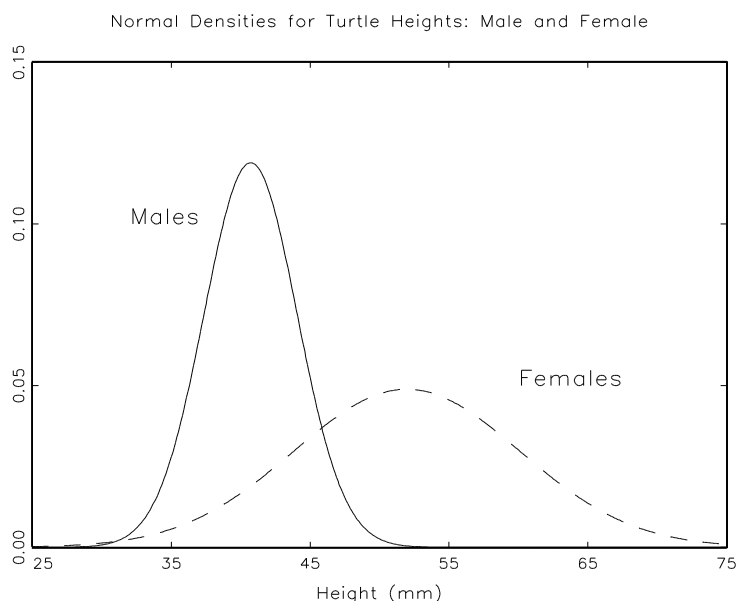


Figure 6: Normal densities for male and female turtle heights.

The cumulative distribution function for the standard normal, i.e.  $P(Z \leq z)$ , is denoted by

$$\Phi(z) = P(Z \leq z).$$

There does not exist a closed form expression for the  $\Phi$  function and numerical integration is needed to compute cumulative probabilities for the standard normal distribution. The cumulative standard normal probabilities have been computed and textbooks usually come equipped with standard normal probability tables. Standard normal probabilities can also be computed using statistical software packages. For example, in SAS, the function

$$x = \text{probnorm}(2);$$

gives the cumulative probability  $\Phi(2) = P(Z \leq 2)$  which is (approximately) equal to 0.97725.

The reason the standard normal distribution is important is because it acts as a benchmark for comparisons. As we shall see later, test statistics are usually standardized differences between an estimated parameter value and a hypothesized parameter value. Also, probabilities for any normal random variable  $X$  can be computed by first “standardizing” the random variable. The following fact demonstrates this:

**FACT:** If  $X$  has a normal distribution with mean  $\mu$  and standard deviation  $\sigma$  then

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

has a standard normal distribution.

If  $X \sim N(\mu, \sigma^2)$ , and we want to find  $P(X \leq a)$ , then we can write

$$\begin{aligned} P(X \leq a) &= P\left(\frac{X - \mu}{\sigma} \leq \frac{a - \mu}{\sigma}\right) \\ &= P\left(Z \leq \frac{a - \mu}{\sigma}\right) \\ &= \Phi\left(\frac{a - \mu}{\sigma}\right) \end{aligned}$$

Also, if we want to compute  $P(a \leq X \leq b)$  for two numbers  $a < b$ , it follows that

$$P(a \leq X \leq b) = \Phi\left(\frac{b - \mu}{\sigma}\right) - \Phi\left(\frac{a - \mu}{\sigma}\right).$$

*In order to compute a probability for a normal random variable  $X \sim N(\mu, \sigma^2)$ , one usually has to standardize it first:*

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

unless the software one is using allows non-standard normal computations.

Returning to the vole example with  $X$  denoting the random variable for skull length, assume  $\mu = 2385.98$  and  $\sigma = 93.32$ . Assuming this distribution is (approximately) normal, we have that

$$Z = \frac{X - \mu}{\sigma} = \frac{X - 2385.98}{93.32}$$

has a standard normal distribution. If we want to compute the probability that a skull length exceeds 2600, then we can write

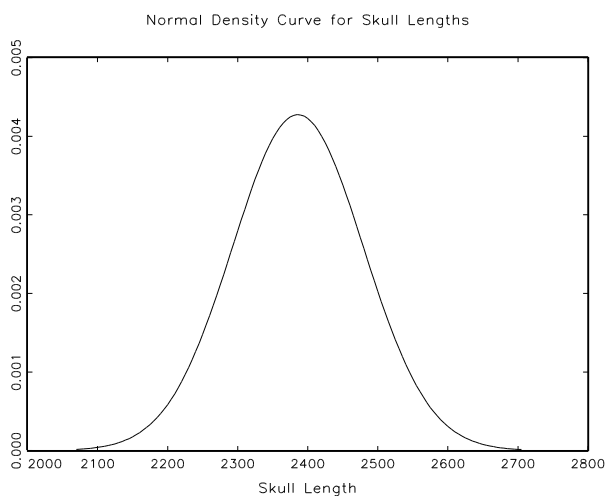
$$\begin{aligned} P(X > 2600) &= 1 - P(X \leq 2600) \quad (\text{Law of Complement}) \\ &= 1 - P\left(\frac{X - \mu}{\sigma} \leq \frac{2600 - \mu}{\sigma}\right) \\ &= 1 - P\left(Z \leq \frac{2600 - 2385.98}{93.32}\right) \\ &= 1 - \Phi(2.2934) \\ &= 1 - 0.98909 \quad (\text{from SAS's probnorm function}) \\ &= 0.010913. \end{aligned}$$

According to the normal probability model, only about 1% of this species of voles have skull lengths exceeding 2600 units.

**Class Exercise:** To further illustrate normal probability computations, let us compute

$$P(2200 < X < 2500)$$

in the vole example and shade the area under the normal pdf in Figure 2 below corresponding to this event. Use SAS's probnorm function (or some other statistical software program) to compute this probability.



Normal density curve with  $\mu = 2386$  and  $\sigma = 93.3$ .

**Relation between probabilities and  $\sigma$ .** The empirical rule for mound shapes distributions holds for the normal distribution. In particular, for any normal random variable  $X$ , the probability that  $X$  lies within

- One standard deviation of its mean is approximately 68%;
- The probability that  $X$  lies within 2 standard deviations of its mean is roughly 95%;
- and the probability that  $X$  lies within 3 standard deviations of its mean is 99.7% approximately.

This empirical rule is often handy to get a quick estimate of a probability.

## Normal Percentiles

In many applications of continuous distributions, the problem is not to compute a probability, but to go in the opposite direction: finding the value of the random variable that leads to a particular probability. For instance, if you take a young child for a doctor's visit and you are told that your child is in the 90th percentile for weight, that means that your child weighs more than 90% of the other children at that age.

In order to demonstrate percentiles, we will begin with the standard normal distribution. Suppose  $Z$  is a standard normal random variable and we want to find the 90th percentile of  $Z$ . That is, we want to find a value, call it  $z_{0.9}$ , so that

$$P(Z \leq z_{0.9}) = 0.9.$$

Graphically, we want to find the value of  $z$  on the horizontal axis so that the area under the standard normal density curve to the left of  $z$  is equal to 0.9. In the above equation, we are given the probability (0.9) and we have to find the corresponding  $z$  value. This can be done in SAS using the *probit* function. The *probit* function in SAS is the standard normal inverse function. Evaluating this function at  $p$  (*probit*( $p$ )) will find the  $z$  value so that the area under the normal density curve to the left of  $z$  is  $p$  where  $p$  is any number between 0 and 1. Here is a SAS example for finding the 90th percentile for the standard normal distribution (i.e.  $p = 0.9$ ).

```
data;
z = probit(.9);
proc print;
run;
```

This gives the 90th percentile for the standard normal which is approximately 1.28155.

In the next chapter we will introduce *confidence intervals* which are used to estimate parameters. A common problem in confidence interval estimation is to find the  $z$ -value so that 95% of the area under the standard normal density lies between  $\pm z$ . If 0.95 of the area lies in between  $\pm z$ , then that leaves the remaining 0.05 for the right and left tails of the distribution, or  $0.05/2 = 0.025$  in each tail. That means we need to find the value  $z_{0.025}$  so that

$$0.025 = \Phi(-z_{0.025}).$$

From SAS's probit function, we find  $z_{0.025} = -1.96$ . Thus,

$$P(-1.96 \leq Z \leq 1.96) = 0.95.$$

In practice, we need to find percentiles for non-standard normal random distributions.

**Example.** Based on a study of body fat percentages of adult men, suppose the average fat percentage is  $\mu = 18.9\%$  and the standard deviation is  $\sigma = 7.71$ . Also assume that body fat percentages follow an approximate normal distribution. For symmetric distributions, the 50th percentile equals the mean value  $\mu$  which in this case is 18.9%. What is the 90th percentile of fat percentages for adult men? Let  $X$  denote a random variable for the body fat percentage of a randomly chosen adult male. Let us call the 90th percentile  $x$ . Then we want to find the value of  $x$  so that  $P(X \leq x) = 0.90$ . If we standardize  $X$ , we get  $P(Z \leq \frac{x-\mu}{\sigma}) = 0.90$  which is the same as  $\Phi(\frac{x-\mu}{\sigma}) = 0.90$ . This implies that  $\frac{x-\mu}{\sigma} = z_{0.9}$ . Solving for  $x$  gives

$$x = \mu + z_{0.9}\sigma.$$

In the previous example we found that  $z_{0.9} \approx 1.28$ . Thus, the 90th percentile for the bodyfat distribution is

$$\mu + z_{0.9}\sigma = 18.9 + 1.28(7.71) = 28.77\%.$$

**z-score.** It is useful to express values on a standardized scale. In the previous body fat example, suppose an adult male has a body fat percentage of 32%. How does he compare to the population of adult males? We can express his body fat percentage in terms of a  $z$ -score, which is the standardized variable:

$$\mathbf{z\text{-score:}} \quad z = \frac{x - \mu}{\sigma}.$$

For the man whose body fat percentage is 32%, his  $z$ -score is

$$z = \frac{32 - \mu}{\sigma} = \frac{32 - 18.9}{7.71} = 1.7.$$

This man's body fat is 1.7 standard deviations above the average value. According to the empirical rule, observations corresponding to  $z$ -scores outside the range of  $\pm 3$  are quite extreme. Of the  $n = 252$  observations in this study, the lowest fat percentage recorded is 0 and the highest was 45.1. Thus, only one observation (the highest) is more than three standard deviations beyond the average value.

### 3 Other Continuous Distributions

The most important continuous distribution is the normal distribution. A couple reasons for this are:

1. Data collected on particular variables often exhibit approximately normal distributions indicated by a bell-shaped histogram.
2. Many statistics are computed by summing observations and due to the *central limit theorem* (see next chapter), sums of random variables tend to behave like normal random variables.

Nonetheless, there are many examples of data sets where the measured variable follows a non-normal distribution. We commented on some of these examples in Chapter 2 when we described the shape of various distributions. For example, a distribution may be mixture of two sub-distributions resulting in a bimodal density, which is clearly non-normal.

Much of classical statistical analysis is based on the assumption that the data is from a normal population. Often practitioners have made the normality assumption without checking the validity of the assumption. It is good statistical practice to assess the normality of the data if the statistical inference techniques are sensitive to this assumption.

In this section we discuss some well-known **non-normal** continuous distributions:

- **Uniform Distribution.** The density function for the uniform distribution is constant over an interval indicating that the probability is spread out uniformly over the interval (instead of concentrated about the mean).
- **Gamma Distribution.** Like the normal distribution, the Gamma Distribution is parameterized by two parameters known as shape and scale parameters respectively. The Gamma distributions are skewed to the right and only take positive values. By varying the values of the two parameters, the Gamma distribution is useful for modeling many skewed right distributions.
- **Exponential Distribution.** This is actually a special case of the Gamma Distribution whose density function is given by

$$f(x; \theta) = (1/\theta)e^{-x/\theta}, \quad x > 0,$$

where  $\theta > 0$  is the parameter. Data was collected on the amount of rain per rainfall in Allen County, Ohio. Figure 7 shows a histogram of the data and overlaid is the exponential density curve which appears to give a fairly good fit to the data.

**Question:** Why would the rainfall distribution be skewed to the right?

- **Log-normal Distribution.** A random variable  $X$  is said to have a *log-normal* distribution if the natural logarithm of  $X$  has a normal distribution:

$$Y = \ln(X) \sim N(\mu, \sigma^2).$$

There are many biological examples of data that are skewed to the right which can be modeled quite well by the log-normal distribution.

There are many other continuous distributions, but the ones mentioned here are some of the most well-known and useful.

Rain Data and Exponential Fit

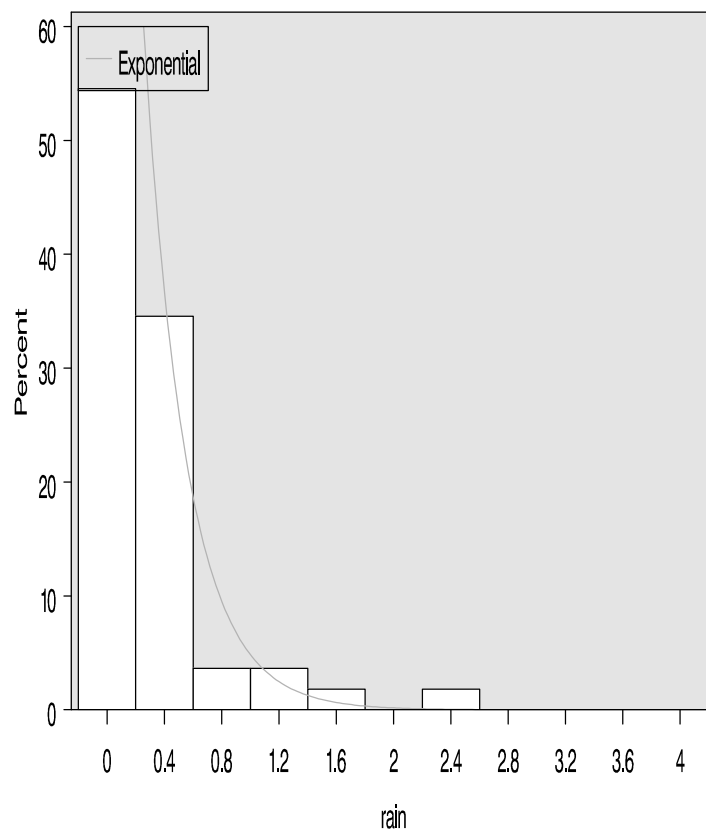


Figure 7: Histogram of rainfall data from Allen County, Ohio where the amount of rain was recorded in inches (to the nearest  $1/4$  inch). Overlaid is an exponential density function.

## Linear Combinations of Random Variables

It is quite common to work with linear combinations of measurements in practice. Suppose  $X_1$  and  $X_2$  are random variables and we form a new random variable  $Y = a_0 + a_1X_1 + a_2X_2$  where  $a_0, a_1$ , and  $a_2$  are constants. If our interest lies in  $Y$ , then we need to know how  $Y$  behaves. That is, we need to have some idea of the probability distribution for  $Y$ . Below are several common examples of linear combinations of random variables that are used frequently in practice.

**Definition.** Given a set of random variables  $X_1, X_2, \dots, X_n$ , a **linear combination** is of the form:

$$L = c_1X_1 + c_2X_2 + \dots + c_nX_n,$$

where  $c_1, c_2, \dots, c_n$  are constants.

Some common examples of linear combinations:

- *Sample Mean:*  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i = \frac{1}{n}X_1 + \frac{1}{n}X_2 + \dots + \frac{1}{n}X_n$ . In this example,  $c_i = 1/n$  for  $i = 1, 2, \dots, n$ .
- *A Difference:*  $X_1 - X_2$ . Here  $c_1 = 1$  and  $c_2 = -1$ .
- *A Contrast:*  $X_1 - (X_2 + X_3)/2$ . Here  $c_1 = 1$  and  $c_2 = c_3 = -1/2$ . Note that the coefficients add to zero in this example:  $c_1 + c_2 + c_3 = 1 - 1/2 - 1/2 = 0$ . This contrast can be thought of as a comparison of the first measurement  $X_1$  with the average of the second and third measurements:  $(X_2 + X_3)/2$ .

Next we give some mathematical results for linear combinations of random variables.

**FACT 1.** The expected value of a linear combination of random variables is equal to the linear combination of the expected values:

$$E[c_1X_1 + c_2X_2 + \dots + c_nX_n] = c_1E[X_1] + c_2E[X_2] + \dots + c_nE[X_n].$$

One of the consequences of this fact is that if  $X_1, \dots, X_n$  represent a random sample from a population with mean  $\mu = E[X_i]$ , then

$$E[\bar{X}] = \mu,$$

the sample mean is *unbiased* for the population mean. That is, the sample mean does not systematically over nor under-estimate the true population mean.

If we form a linear combination of random variables, then Fact 1 tells us the mean of this new random variable. We also need to know the variance for the new random variable. Recall that the variance is the average squared deviation from the mean. For a single random variable  $X$ , we can easily show:

$$\text{Var}(cX) = c^2\text{Var}(X),$$

for any constant  $c$ . If we have a linear combination of *independent* random variables, then we have the following fact:

**FACT 2.** If  $X_1, X_2, \dots, X_n$  are *independent* random variables, then for any linear combination  $c_1X_1 + c_2X_2 + \dots + c_nX_n$ , we have

$$\text{Var}(c_1X_1 + c_2X_2 + \dots + c_nX_n) = c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + \dots + c_n^2\text{Var}(X_n).$$

Note that the assumption of independence is critical here. If the observations are not independent, then the formula generally will not hold. If we obtain a random sample of measurements  $X_1, \dots, X_n$ , then it is assumed that the random variables are independent and therefore the above variance formula holds for a linear combination of the random variables.

A very important consequence of Fact 2 is that it allows us to compute the variance of the sample mean from a random sample. Suppose  $X_1, \dots, X_n$ , denotes a random sample from a population with mean  $\mu$  and variance  $\sigma^2$ . Then, for each  $i = 1, \dots, n$ ,  $E[X_i] = \mu$  and  $\text{Var}(X_i) = \sigma^2$ . If we use the sample mean to estimate  $\mu$  (since, as we saw above, the sample mean is unbiased for  $\mu$ ), then we need to know how  $\bar{X}$  behaves as a random variable. In particular, what is the variance of  $\bar{X}$ ? The answer can be derived from Fact 2:

$$\begin{aligned} \text{Var}(\bar{X}) &= \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right) \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) \quad (\text{From Fact 2}) \\ &= \frac{1}{n^2} [\sigma^2 + \sigma^2 + \dots + \sigma^2] \quad (n\text{-times}) \\ &= \sigma^2/n. \end{aligned}$$

The variance of a single random variable is  $\sigma^2$ , but the variance of the sample mean is smaller by a factor of  $1/n$ . As the sample size  $n$  gets larger, the variance of the sample mean gets smaller. That is one advantage of a large sample size – very precise estimation of the population mean.

If we have a collection of independent random variables  $X_1, \dots, X_n$ , the previous two facts tell us the mean and variance of any linear combination of these random variables. However, the two facts do not tell us what sort of distribution the linear combination will have. Typically, linear combinations of random variables can have very complicated probability distributions. However, if the random variables are normally distributed, then it follows that any linear combination of the random variables will also have a normal distribution:

**FACT 3.** If  $X_1, \dots, X_n$  are independent *normal* random variables with means  $\mu_1, \dots, \mu_n$  and variances  $\sigma_1^2, \dots, \sigma_n^2$  respectively, then the distribution of any linear combination

$$c_1X_1 + c_2X_2 + \dots + c_nX_n,$$

will have a normal distribution with mean  $c_1\mu_1 + \dots + c_n\mu_n$  and variance  $c_1^2\sigma_1^2 + \dots + c_n^2\sigma_n^2$ .

In particular, if  $X_1, \dots, X_n$  is a random sample from a normal distribution  $N(\mu, \sigma^2)$ , then the sample mean also has a normal distribution. Putting all three facts together gives:

$$\bar{X} \sim N(\mu, \sigma^2/n),$$

when sampling from a normal distribution.

### Dependent Random Variables Covariance and Correlation.

In Fact 2 and 3, we assumed that the random variables are independent. This assumption is often false in many interesting examples, in particular, in the realm of regression analysis and multivariate statistics. In regression analysis, we explore the relation between a response variable that is related to one or more “independent” variables. Because these variables are related, they are not independent.

In multivariate statistics, several variables are measured on a single subject or object. For instance, in studying plant growth, we may measure the length and width of the leaves on the plant. Longer leaves will tend to have longer widths, and therefore these two variables will not be independent. It is quite common practice to analyze data where numerous variables have been measured; in such cases, the statistical analysis deals with exploring and quantifying the relationships and dependencies between the variables.

**Covariance.** In order to quantify the relation between two random variables, the *covariance* is used. If we have two random variables  $X$  and  $Y$  that vary jointly, with means  $\mu_x$  and  $\mu_y$ , then we can define a new random variable  $(X - \mu_x)(Y - \mu_y)$ . The covariance between  $X$  and  $Y$  is defined to be the expected value of this new random variable:

$$E[(X - \mu_x)(Y - \mu_y)].$$

A shortcut formula for computing the covariance is:

$$\text{Cov}(X, Y) = E[XY] - \mu_x\mu_y.$$

**Question:** What is the rationale for the covariance and how does it provide information on how  $X$  and  $Y$  *co-vary*? If  $Y$  tends to take above average values when  $X$  takes above average values, then the deviations  $(X - \mu_x)$  and  $(Y - \mu_y)$  will both be positive and their product will be positive. If  $Y$  tends to take below average values when  $X$  takes below average values, then  $(X - \mu_x)$  and  $(Y - \mu_y)$  will both be negative and their product will be positive on average. Therefore, if  $X$  and  $Y$  vary in the same fashion, the covariance will be positive. On the other hand, if  $Y$  tends to be large when  $X$  is small (or vice-versa –  $Y$  tends small when  $X$  is large), then the covariance will be negative.

**Question:** Give an example of two random variables  $X$  and  $Y$  that vary jointly that will have a positive covariance? A negative covariance?

Another measure of association between two random variables that is used more commonly is the *correlation*, which can be thought of as the covariance scaled by the respective standard deviations.

**Definition.** The **correlation**, denoted  $\rho$  (“rho”), between two random variables  $X$  and  $Y$  with standard deviations  $\sigma_x$  and  $\sigma_y$  respectively is defined to be

$$\text{Population Correlation: } \rho = \frac{\text{Cov}(X, Y)}{\sigma_x\sigma_y}.$$

The correlation, like the covariance, is a population parameter and must be estimated in practice. If we have a random sample of measurements on  $X$  and  $Y$ :  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ , then the sample correlation, denoted by  $r$  is defined to be the sample counterpart of  $\rho$ :

$$\text{Sample Correlation: } r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

Here are some facts about correlations ( $r$  and  $\rho$ ):

1. Correlations are always between  $-1$  and  $1$ :

$$-1 \leq \rho \leq 1.$$

2. Correlation is a dimensionless quantity. Random variables are often measured on some scale (grams, centimeters, Fahrenheit, etc.) and therefore means, variances, standard deviations, and covariances also have scales associated with them. However, by the way correlation is defined, it is a scaleless or dimensionless quantity.
3. If the correlation between two random variables  $X$  and  $Y$  is perfect, i.e.  $\rho = \pm 1$ , then  $Y = a + bX$ , for some constants  $a, b$  with  $b \neq 0$ .
4. Correlation is a measure of the strength of the *linear* relationship between two variables. If  $\rho \approx 1$ , then there is a very strong positive relationship. If  $\rho \approx -1$ , then there is a very strong negative relationship. If  $\rho \approx 0$ , then there is a no linear relation or a very weak linear relation between the two variables. Figure 8 shows plots of  $(X, Y)$  data with different correlations. The distribution for the top-left panel had a correlation of  $\rho = 0.95$ . The plot shows a strong positive relation between  $X$  and  $Y$  with the points tightly clustered together in a linear pattern. The correlation for the top-right panel is also positive with  $\rho = 0.50$  and again we see a positive relation between the two variables, but not as strong as in the top-left panel. The bottom-left panel corresponds to a correlation of  $\rho = 0$  and consequently, we see no relationship evident between  $X$  and  $Y$  in this plot. Finally, the bottom-right panel shows a negative linear relation with a correlation of  $\rho = -0.50$ .
5. If the correlation between two variables is high, this does not necessarily mean that one variable has a *causal* relation to the other variable.
  - Let  $X$  equal the amount of phosphorus in the soil and  $Y$  equal the height of a plant. It makes sense in this example that higher phosphorus levels will cause the plant to grow higher (provided there is not too much phosphorus) and a positive correlation is expected.
  - A survey of U.S. cities is conducted. For each, record  $X =$  number of people who attend church weekly and  $Y =$  number of murders. Guess what – these two variables are positively correlated. Does having a large number of people going to church cause murder rates to go up? The answer of course is no. Both of these variables are related to the overall population of the cities. Cities with large populations will tend to have large numbers of people attending church weekly simply because they are large cities. Large cities will also tend to have more murders than smaller cities again because they have more people than small cities.
6. Two variables may be related, but the relation may be nonlinear in which case the correlation is not an appropriate measure of association. Remember: correlation is a measure of *linear* association between two variables. In the phosphorus example above, if too much phosphorus is in the soil, it will have a detrimental effect on the plant leading to lower plant heights. Figure 9 shows a plot of data one might expect to see in this example. There is clearly a very strong relation between  $X$  and  $Y$ , but the relation is nonlinear. The sample correlation is nearly zero but it would be wrong to say the two variables are unrelated simply because the correlation is near zero.  $X$  and  $Y$  are strongly associated, but not in a linear fashion.
7. Just because the correlation between two variables is high does not necessarily mean that the two variables are linearly related. Two variables with a slight nonlinear association can produce high correlations. It is always a good idea to plot your data to see what sort of relation exists between variables.

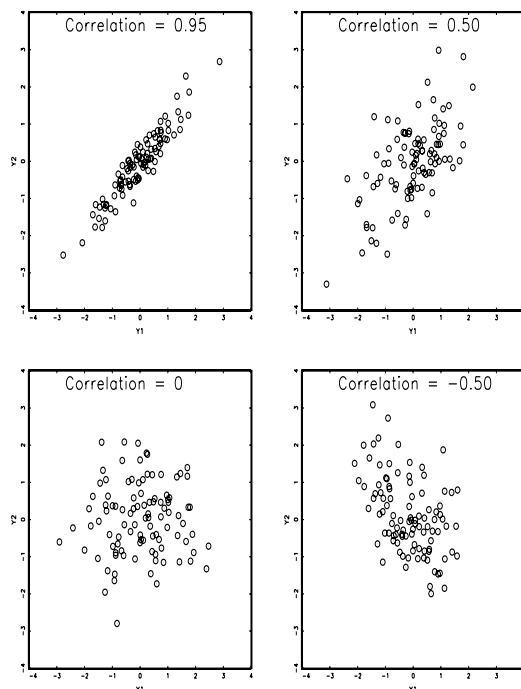


Figure 8: Scatterplots of data obtained from bivariate distributions with different correlations.

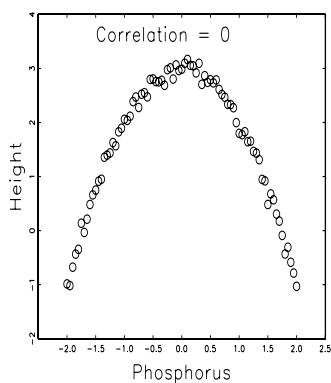


Figure 9: A scatterplot showing a very strong but nonlinear relationship between  $y_1$  and  $y_2$ . The correlation is nearly zero.

Returning to our discussion of linear combinations of variables, we have the following general result:

**FACT 4:** For a linear combination of two dependent random variables  $X_1$  and  $X_2$  we have:

$$\text{Var}(c_1X_1 + c_2X_2) = c_1^2\text{Var}(X_1) + c_2^2\text{Var}(X_2) + 2c_1c_2\text{Cov}(X_1, X_2).$$

This formula easily generalizes to any number of dependent random variables. Note that if  $X_1$  and  $X_2$  are independent, then the covariance between them is zero, and the formula for Fact 4 is identical to the formula for Fact 2 when  $n = 2$ .

We now present our final fact regarding correlation and linear combinations:

**FACT 5:** Suppose the correlation between  $X$  and  $Y$  is  $\rho$  and we linearly transform  $X^* = a_1 + b_1X$  and  $Y^* = a_2 + b_2Y$ . If  $b_1$  and  $b_2$  have the same sign, then the correlation between  $X^*$  and  $Y^*$  is still equal to  $\rho$ ; if  $b_1$  and  $b_2$  have opposite signs, then the correlation between  $X^*$  and  $Y^*$  will be  $-\rho$ . In other words, a linear transformation of variables does not change the correlation (except perhaps up to a sign change).

Here is an illustration with sample data  $(x_1, y_1), \dots, (x_n, y_n)$ . Let  $x_i^* = a_1 + b_1x_i$  and  $y_i^* = a_2 + b_2y_i$ . Assume for illustration that  $b_1, b_2 > 0$ . Then from our previous facts on linear combinations, we have:  $\bar{x}^* = a_1 + b_1\bar{x}$ ,  $\bar{y}^* = a_2 + b_2\bar{y}$ . Also the standard deviation  $s_{x^*}$  of the  $x^*$  is  $b_1s_x$ . Similarly,  $s_{y^*} = b_2s_y$ . Then the sample correlation between  $x^*$  and  $y^*$  from the correlation formula is

$$\begin{aligned} \frac{\sum (x_i^* - \bar{x}^*)(y_i^* - \bar{y}^*)/(n-1)}{s_{x^*}s_{y^*}} &= \frac{\sum (a_1 + b_1x_i - a_1 - b_1\bar{x})(a_2 + b_2y_i - a_2 - b_2\bar{y})/(n-1)}{b_1s_x b_2s_y} \\ &= \frac{\sum b_1b_2(x_i - \bar{x})(y_i - \bar{y})/(n-1)}{b_1s_x b_2s_y} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})/(n-1)}{s_x s_y} \\ &= \text{Correlation between } x \text{ and } y. \end{aligned}$$

For example, suppose the correlation between the length and width of a leaf in centimeters is  $r = 0.86$ . If we convert the length and width to inches by multiplying each by 2.54, the correlation between the length and width in inches is still 0.86.

### Normal Approximation to the Binomial Distribution.

Recall that the binomial distribution (i.e. number of successes out of  $n$  identical and independent trials where each trial results in a success or failure) is a discrete distribution. Figure 10 shows the probability distribution function for the binomial distribution when  $n = 10$  (left panel) and  $n = 100$  (right panel) and the success probability is  $= 0.8$  in both cases. When  $n = 10$ , the distribution is skewed somewhat to the left. However, when the number of trials is large ( $n = 100$  in the right panel), the distribution looks very much like the normal bell-shaped curve.

If  $n$ , the number of trials in a binomial experiment, is large and the success probability  $p$  is not too close to either zero or one, then the binomial distribution can be well approximated by the continuous normal distribution. This is a consequence of the *central limit theorem* which we shall discuss in the next chapter.

Recall that the mean and variance of a binomial random variable  $X$  with  $n$  trials and success probability  $p$  are  $\mu = np$  and  $\sigma^2 = np(1 - p)$  respectively. If  $n$  is sufficiently large, then

$$Z = \frac{X - np}{\sqrt{np(1 - p)}}$$

will follow an approximate standard normal distribution. In particular, we can approximate the probability  $P(X \leq a)$  for a constant  $a$  by the following:

$$\begin{aligned} P(X \leq a) &= P\left(\frac{X - np}{\sqrt{npq}} \leq \frac{a - np}{\sqrt{npq}}\right) \quad (\text{where } q = 1 - p) \\ &\approx P\left(Z \leq \frac{a + 0.5 - np}{\sqrt{npq}}\right) \\ &= \Phi\left(\frac{a + 0.5 - np}{\sqrt{npq}}\right). \end{aligned}$$

The value of 0.5 above is called the *continuity correction* which improves the normal approximation.

The quality of the normal approximation improves as  $n$  gets larger and/or  $p$  gets closer to  $1/2$ . Note that the binomial distribution is perfectly symmetric and mound shaped when  $p = 1/2$  and it becomes more and more skew as  $p$  gets closer to zero or one. A general rule of thumb is that the normal approximation will be pretty good if  $npq$  is at least 5.

**Example.** Suppose  $n = 25$  subjects with a certain form of cancer take a chemotherapy treatment. The probability of a successful treatment (i.e. remission) for an individual subject is  $p = 0.3$ . What is the probability that the number of successful treatments out of the 25 subjects is at least 10?

Let  $X$  denote the number of successful treatments. Then  $X$  has a binomial distribution with  $n = 25$  and  $p = 0.3$ . Also,  $npq = 25(.3)(.7) = 5.25$  indicating that the distribution of  $X$  should be well approximated by a normal distribution. We want to compute  $P(X \geq 10)$ . We would expect to see  $E[X] = np = 25(0.3) = 7.5$  successes with a variance of  $\sigma^2 = np(1 - p) = 25(0.3)(0.7) = 5.25$ . Using the law of complements, we have

$$P(X \geq 10) = 1 - P(X < 10) = 1 - P(X \leq 9).$$

The exact answer using

```
probbnml(0.3, 25, 9);
```

in SAS gives the

$$1 - P(X \leq 9) = 1 - 0.81056 = 0.18944.$$

Using the normal approximation, we find that

$$1 - P(X \leq 9) = 1 - \Phi\left(\frac{9 + 0.5 - 7.5}{\sqrt{5.25}}\right) = 1 - \Phi(0.8729) = 1 - 0.8086 = 0.1914,$$

which is fairly close to the exact value of 0.18944. Note that if we had not added the continuity correction of 0.5, the normal approximation would be quite poor. Thus, there is only about a 19% chance that 10 or more of the patients will have a successful treatment.

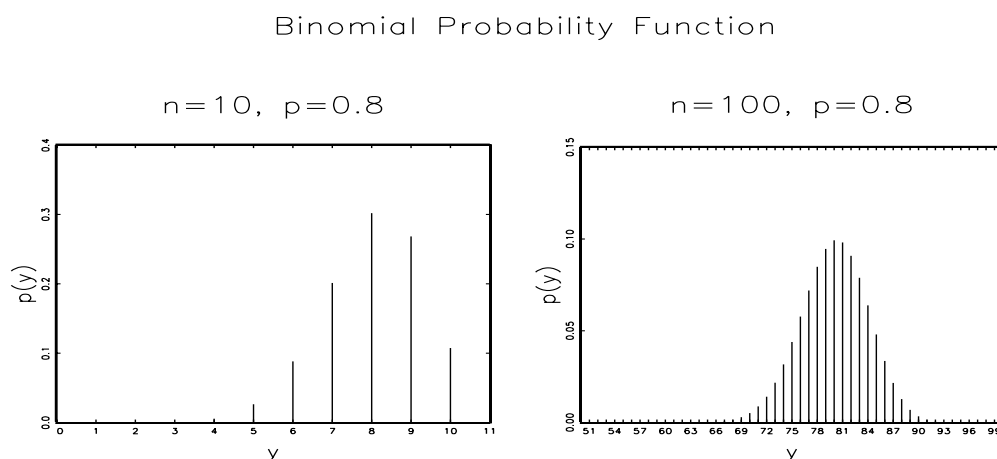
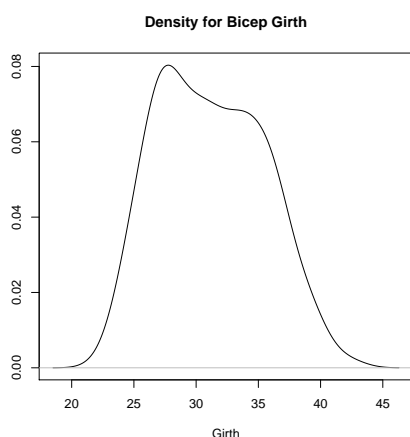


Figure 10: **Binomial Probability Distribution Function** Left Panel:  $n = 10, p = 0.80$ , Right Panel:  $n = 100, p = 0.80$ .

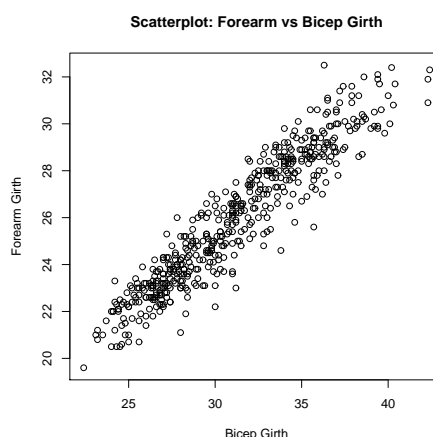
## 4 Problems

Figure (a) below shows the probability density function (pdf) for bicep girth (in centimeters) for the population of adults in the United States. Figure (b) shows a scatterplot of forearm girth versus bicep girth for a sample of  $n = 507$  adults. Use these plots to answer the following questions.

1. Is the distribution for bicep girth normal? Yes or No (circle one)
2. Which of the following best describes the shape of the bicep girth distribution? (Circle one choice)
  - (a) bell-shaped
  - (b) skewed left
  - (c) skewed right
  - (d) bimodal
  - (e) binomial
3. (3 points) Can you think of an explanation for why the pdf in figure (a) has the shape it has?
4. Which of the following is the mean  $\mu$  bicep girth? (Circle one choice)
  - (a) 31
  - (b) 25
  - (c) 40
  - (d) 36
  - (e)  $\bar{x}$
5. (3 points) What proportion of adults have a bicep girth exceeding 35 cm? (Circle one choice):
  - (a) 0.01
  - (b) 0.05
  - (c) 0.22
  - (d) 0.50
  - (e) 0.90
6. Which of the following is the correlation between bicep and forearm girth? (Circle one choice)
  - (a) -1
  - (b) -0.5
  - (c) 0
  - (d) 0.25
  - (e) 0.94
  - (f) 1.0
  - (g) 10.3
7. If we convert the bicep and forearm measurements from centimeters to inches by dividing each measurement by 2.54, what will happen to the correlation between bicep and forearm girth?



(a) Bicep Girth pdf



(b) Scatterplot of Forearm vs Bicep Girth

8. Let  $X$  denote the forearm girth of a randomly selected adult and suppose  $P(X > 25) = 0.6$ . Which of the following equals  $P(X > 28)$ ? (Only one answer makes sense – circle it)
- (a) 0.63      (b) 0.78      (c) 32      (d) 0.27      (e) 0
9. In a morphometric study of plants, data was collected on the width and length of leaves in centimeters. The data revealed a correlation of  $r = 0.58$  between the width and length of the leaves. Suppose the data is transformed to units of inches by multiplying each width and length measurement by 2.54. What will be the correlation between leaf width and length in inches?  $r =$
- (a) 0      (b) 1.4732      (c)  $|2.54| \cdot (0.58)$       (d) 0.58      (e)  $\sqrt{2.54} \cdot (0.58)$
- (f)  $r$  cannot be computed without knowing the data values.
10. The log-concentration of PCB (in ppm) in pelicans follows a normal distribution with mean  $\mu = 5.28$  and standard deviation  $\sigma = 0.4$ . Use this information to answer the following questions:
- What proportion of pelicans have a log-PCB level exceeding 5.4?
  - What proportion of pelicans have a log-PCB between 5 and 6?
  - Find the 95th percentile of the log-PCB concentration for the pelican population.
  - There is a concern that the PCB exposure for pelicans has increased due to the dumping of industrial waste in recent years. A sample of  $n = 50$  pelicans are collected and tested and the sample mean log-PCB for these fifty pelicans was found to be  $\bar{x} = 5.4$ . If we assume the mean and standard deviation for the log-PCB is still 5.28 and 0.4 respectively, what is the probability the sample mean would take a value of 5.4 or greater?
11. Cholesterol levels for men between the ages of 20 to 30 follow a normal distribution with mean  $\mu = 170(mg/dL)$  and standard deviation  $\sigma = 20$ .
- What proportion of men in this age group have cholesterol levels exceeding  $200mg/dL$ ?
  - The cholesterol levels of a random sample of  $n = 20$  men in this age group were measured. Find the probability that the average of these twenty measurements exceeds 200.
  - What is the 95th percentile for the cholesterol distribution? That is, what is the cholesterol level such that only 5% of the men (between 20 and 30) have a cholesterol readings exceeding this level?

12. A study on the size of voles and their offspring was conducted. Let  $X_1$  equal the height of a randomly selected mother and  $X_2$  equal the height of the mother's daughter and suppose that  $\text{var}(X_1) = \text{var}(X_2) = 3$ . In studying differences in heights between mothers and their daughters, it was found that  $\text{var}(X_1 - X_2) = 2$ . What is the correlation between  $X_1$  and  $X_2$ ?
13. The log-concentration of PCB (in ppm) in pelicans follows a normal distribution with mean  $\mu = 5.28$  and standard deviation  $\sigma = 0.4$ . Which of the following is the 95th percentile of the log-concentration distribution? (Circle one)
- (a) 4.48      (b) 4.88      (c) 5.28      (d) 5.016      (e) 5.94      (f) 9.08
13. Bioconcentration factors (BCF) represents the equilibrium concentration of a toxicant in an organism. Suppose the toxicant under consideration is PCB's and that the average value of the BCF in snails at a particular site is  $\mu = 100$  with standard deviation  $\sigma = 8$ . Suppose further the the distribution of BCF's among snails at the site varies according to a normal distribution.
- a) Find the probability that a snail selected at random has a BCF exceeding 115.
- b) What proportion of snails have BCF's between 90 and 100?
- c) What is the 75th percentile of the PCB concentration for snails at the lake?
- d) Suppose 10 snails are selected at random. What is the probability that 8 of the 10 snails have a BCF exceeding 115? (Hint: The solution to part (a) is useful here).
14. Body temperatures for individuals fluctuate during the day. In a hospital, nurses measure the body temperatures of patients in the morning and in the evening. The correlation between the morning and evening body temperatures readings is very high. Which of the following is a reasonable correlation between the morning and evening temperature readings? (Circle One)
- a) 0.93    b) -0.93    c) 0.001    d) 98.6    e) 98.2    f) 99.6    g) 0.50    h) 49.3
15. Let  $X$  be a random variable equal to the spinal bone density of a randomly selected middle-aged woman. The distribution of spinal bone density is normal with mean  $\mu = 0.80$  and standard deviation  $\sigma = 0.13g/cm^2$ . Which of the following probabilities is *largest*? (Circle one)
- (a)  $P(X > 0.93)$       (b)  $P(X < 0.93)$       (c)  $P(X < 0.67)$       (d)  $P(X > 0.80)$       (e)  $P(X > 1.06)$
16. A treatment is available for increasing bone density. Suppose  $X_1$  is the bone density before treatment and  $X_2$  is the bone density after treatment and suppose that  $\text{var}(X_1) = \text{var}(X_2) = 0.0169$ . Suppose further that the correlation between  $X_1$  and  $X_2$  is  $\rho = 0.8$ . Which of the following is  $\text{var}(X_2 - X_1)$ ? (Circle one):
- (a) 0      (b) 0.0338      (c) 0.00676      (d) -1.5662
- (e) Not enough information – we need to know the expected values of  $X_1$  and  $X_2$ .
17. **Normal approximation to the binomial.** Let  $X$  denote a binomial random variable. Let  $Z = \frac{X-\mu}{\sigma}$  denote the standardized version of  $X$  where  $\mu$  and  $\sigma$  are the binomial mean and standard deviation.
- a) If  $n = 20$  and  $p = 0.4$ , compute the exact probability  $P(X \leq 10)$
- b) Compute the approximate probability in (a) by  $P(Z \leq \frac{10.5-\mu}{\sigma})$  using a normal distribution. Are the two probabilities similar to each other?

## References

Airoldi, J. P., Flury, B. and Salvioni, M., (1996), "Discrimination between two species of *Microtus* using both classified and unclassified observations. *Journal of Theoretical Biology*, **177**, p 247–262.