

April 6, 2006

Chapter 3. Continuous Models

In this chapter we introduce models for measurements obtained on a continuous scale.

3.1 Introduction.

Many experiments, the response of interest can take values in a continuum, over an interval of numbers. For instance, measuring the fill-weight of soup dispensed into a cup. Fill-weight is a continuous variable that can theoretically assume values on the positive real line. If we measure the time until a machine breaks down, that too is a measurement on a continuous scale. In this section we discuss *continuous* probability distributions and random variables. The mathematics for continuous distributions has to be different than that for discrete probability distributions because we cannot assign a positive probability to every number on an interval of real numbers – if we did, we would have a total probability of infinity which is not allowed. Instead, for continuous probability distributions, it makes sense to assign non-negative probabilities over intervals of real numbers.

It's important to note that the outcomes of many experiments are continuous responses but we cannot actually measure such outcomes on a truly continuous scale. For instance, we may only be able to measure the fill-weight of soup in a cup to the closest tenth of a milliliter yielding data on a discrete scale. If the precision of the measurements is quite good, then the distinction between the continuous outcome and the discrete measuring scale may be of small importance. However, it may be necessary in some applications to take the discreteness into consideration, particularly if the measuring instrument is not very precise.

In order to help understand the continuous probability model, consider once again the cup-a-soup example. Figure 1 shows a histogram of the fill-weights of soup in a sample of cups from one of the production lanes at the plant. The fill-weights were divided into 10 intervals for plotting the histogram in Figure 1. However, the fill-weights vary continuously, so we can obtain a better picture of the distribution of fill-weights using a finer partition for the histogram, particularly since we have quite a large sample size, as seen in Figure 2. We can approximate this finer histogram by a smooth function, as in Figure 3. Figure 3 shows a rather irregular smooth

curve overlaying the histogram. Recall that there was a problem with the production process because the soup fill-weights were varying too much. In particular, as can be seen in Figure 3, there are a number of cups receiving far too little soup – the distribution looks skewed to the left.

In many other situations, where there are no problems with the population, or if the population is homogeneous, then the smoothed histogram will have a nice regular

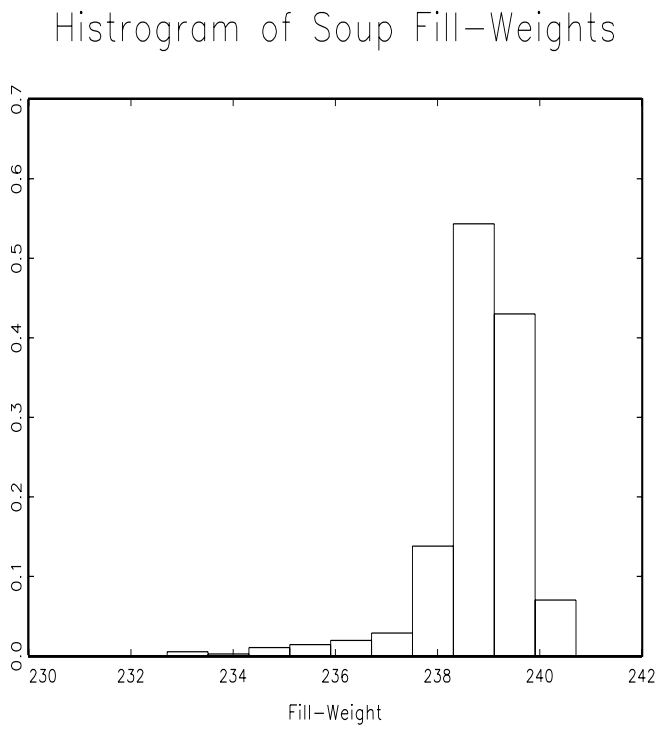


Figure 1: A coarse histogram of fill-weights for lane 1 of the cup-a-soup production sample of $n = 969$ cups.

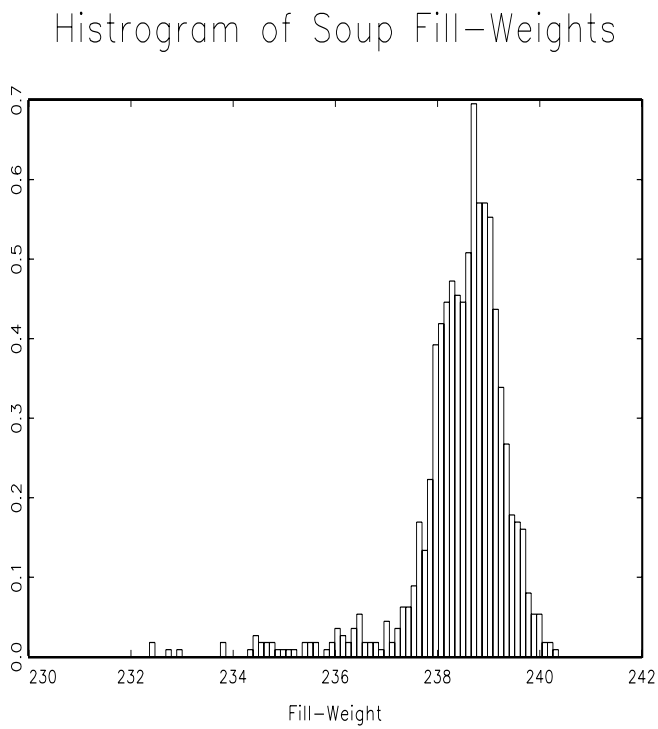


Figure 2: A finer histogram of soup fill-weights.

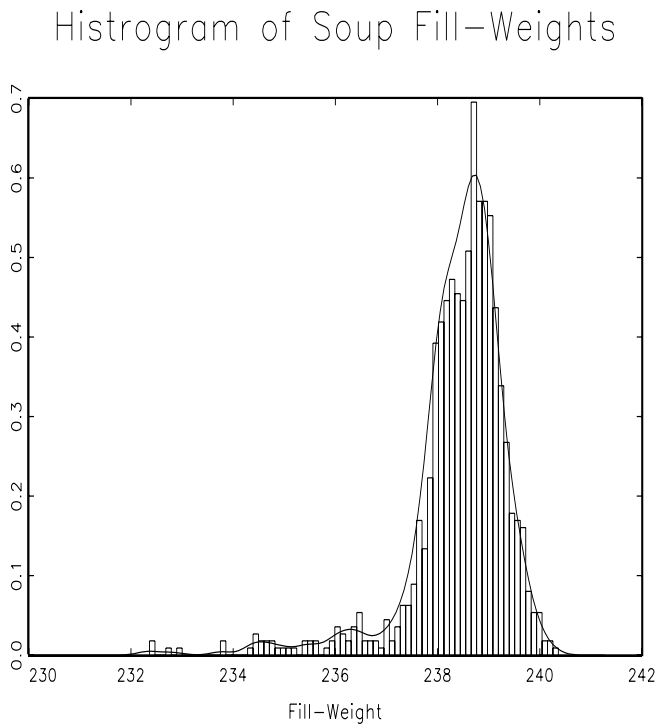


Figure 3: Soup fill-weight histogram with a probability density overlaid.

shape. To illustrate, consider the next example.

Swiss Gas Masks. Several years ago, the Swiss Army began work on designing new gas masks for its soldiers. Anthropologists carefully collected data (Flury 1997) from a sample of Swiss soldiers. The data consisted of six measurements of the head:

- MFB = Minimal frontal breadth (forehead width)
- BAM = Breadth of angulus mandibulae (chin width)
- TFH = True facial height
- LGAN = Length from glabella to apex nasi (tip of nose to top of forehead)
- LTN = length from tracion to nasion (top of nose to ear)
- LTG = Length from tracion to gnathion (bottom of chin to ear)

One of the things to note here is that the mask fitting problem is a multivariate problem. If we select sizes considering only facial height say, the resulting fits may be poor. Incorporating all six variables makes the problem much more complicated.

In order to illustrate continuous probability distributions, we will focus on just one variable, LTG which is roughly a measure of length from the bottom of the chin to the ear. Figure 4 shows a histogram of the LTG measurements from a sample of $n = 200$ Swiss soldiers. The histogram shows roughly a symmetric *bell-shaped* pattern which is quite common in many applications. Note that the histogram is somewhat “uneven” and not completely smooth. This is a consequence of the fact that we have only a tiny sample from a larger population. If we are interested in fitting masks to the

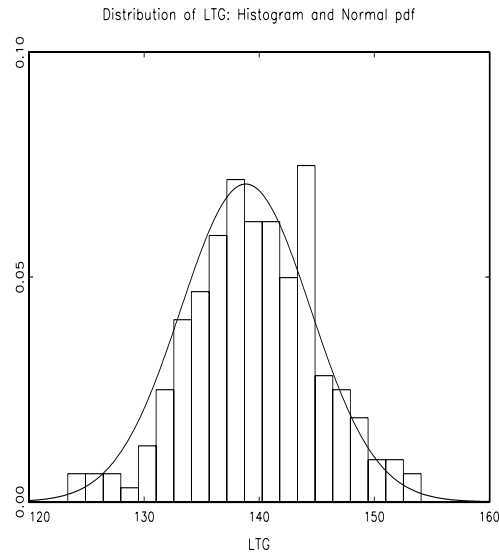


Figure 4: The distribution of the length from tragon to gnathion (LTG) for $n = 200$ Swiss soldiers. The plot shows a histogram and overlaid with this is a normal probability density function (pdf).

entire population of interest (which may include future conscripts), then it would be wise to try to find a probability model for the entire population instead of focusing only on the histogram. Often in practical situations, the more data one obtains, the smoother the resulting histogram. Overlaid on the histogram in Figure 4 is a *normal probability density function (pdf)* which provides a nice approximation to the histogram (see below for formal definitions). The normal pdf depends on two parameters only (μ and σ). Thus, knowing the values of these two parameters tells us everything about the population. One of the goals of statistics is to use the data to estimate the true values of the parameters.

3.2 Probability Density Functions

In practice, the shape of continuous distributions vary considerably depending on the application at hand. However, as mentioned above, the symmetric bell-shaped distribution occurs quite frequently. Now we give a general definition of a probability density function which is used to model the behavior of continuous random variables and allows for the computation of probabilities.

Definition. Let Y denote a continuous random variable. The function $f(y)$ is called the *probability density function (pdf)* for the distribution of Y if

1. $f(y) \geq 0$ for all real numbers y (since probabilities cannot be negative).
2. $\int_{-\infty}^{\infty} f(y)dy = 1$, i.e. the total probability must be one.
3. For any two real numbers $y_1 \leq y_2$ we have

$$P(y_1 \leq Y \leq y_2) = \int_{y_1}^{y_2} f(y)dy.$$

From this definition, it follows that $P(Y = y) = 0$ for any real number y . The reason why is that for continuous random variables, probabilities are computed by integrating the pdf which corresponds to finding the area under the curve for $f(y)$ according to property (3) above. The area under the curve at a single point is zero. Students often find the fact that $P(Y = y) = 0$ for continuous random variables a bit paradoxical. For instance, suppose Y is a continuous random variable representing the height of an adult. Then $P(Y = 6 \text{ feet}) = 0$ which seems to be saying that it is impossible for someone to have height exactly 6 feet. One way to think about the problem is that $P(Y = 6 \text{ feet})$ represents the proportion of people who are exactly 6 feet tall. However, no two people are exactly the same height when measured with unlimited precision. Also, continuous distributions model infinite populations. In practice a population may not be infinite, but the continuous model often provides a very good approximation for very large populations. From this discussion it follows that $P(y_1 \leq Y \leq y_2) = P(y_1 \leq Y < y_2) = P(y_1 < Y < y_2)$.

The cumulative distribution function (cdf) for continuous distributions is defined by

$$F(y_0) = \int_{-\infty}^{y_0} f(y)dy.$$

Thus, $P(Y \leq y_0) = F(y_0)$.

The expected value and variance of a continuous random variable can be found in a similar fashion to what was done for discrete random variables except that we replace the summation symbol by integration for continuous random variables.

The expected value (or mean) of a continuous random variable Y with pdf $f(y)$ is given by

$$\mu = \int_{-\infty}^{\infty} yf(y)dy$$

and the variance of Y is given by

$$\sigma^2 = E[(Y - \mu)^2] = \int_{-\infty}^{\infty} (y - \mu)^2 f(y)dy.$$

The same shortcut formula for variance that held for discrete distributions also holds for continuous distributions, namely

$$\sigma^2 = E[Y^2] - \mu^2 = \int_{-\infty}^{\infty} y^2 f(y)dy - \mu^2.$$

In order to see the mechanics behind the computations, we next consider a simple example.

Example. Suppose the continuous random variable Y has a pdf given by

$$f(y) = 3y^2, \text{ for } 0 < y < 1,$$

and zero otherwise. Find the mean and variance of Y . Also, compute the probability that $0.25 < Y < 0.75$.

The mean or expected value of Y is given by the following computation:

$$\begin{aligned}
 \mu &= \int_{-\infty}^{\infty} yf(y)dy \\
 &= \int_0^1 y\{3y^2\}dy \\
 &= 3 \int_0^1 y^3 dy \\
 &= 3\left(\frac{1}{4}y^4\right)\Big|_0^1 \\
 &= \frac{3}{4}(1^4 - 0^4) \\
 &= 3/4.
 \end{aligned}$$

In order to compute the variance, first we shall compute $E[Y^2]$ by

$$\begin{aligned}
 E[Y^2] &= \int_{-\infty}^{\infty} y^2 f(y) dy \\
 &= 3 \int_0^1 y^4 dy \\
 &= 3/5.
 \end{aligned}$$

Therefore,

$$\begin{aligned}
 \sigma^2 &= E[Y^2] - \mu^2 \\
 &= 3/5 - (3/4)^2 \\
 &= 0.0375.
 \end{aligned}$$

In order to compute the probability that Y lies between 0.25 and 0.75, we simply compute the following integral:

$$\begin{aligned}
 P(0.25 < Y < 0.75) &= \int_{0.25}^{0.75} 3y^2 dy \\
 &= y^3 \Big|_{0.25}^{0.75} \\
 &= 0.40625.
 \end{aligned}$$

This probability is indicated graphically in Figure 6 where the probability corresponds to the shaded region in Figure 6.

3.3 Some Important Continuous Distributions.

Many of the data sets observed in practice produce histograms with distinct shapes which can be well approximated by given probability models. The most important

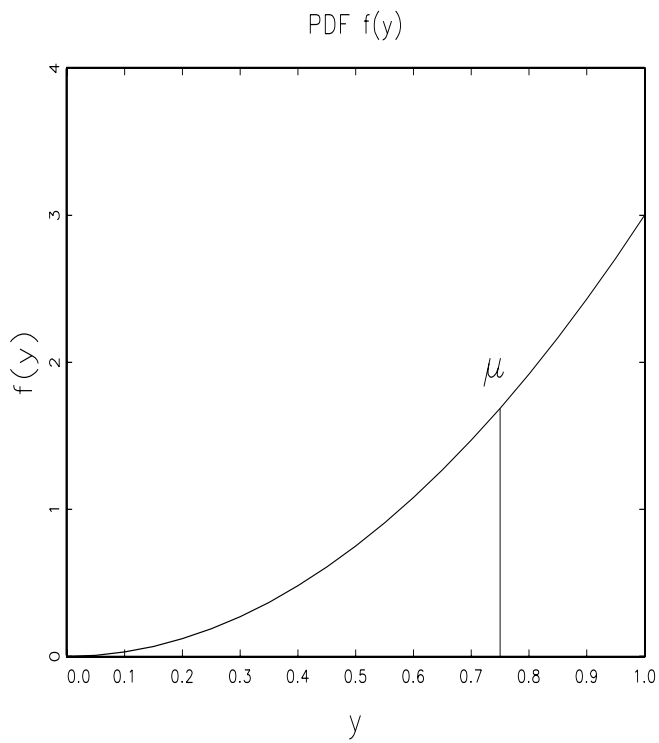


Figure 5: The plot of the pdf $f(y) = 3y^2$, for $0 < y < 1$. The vertical line indicates the mean of the distribution.

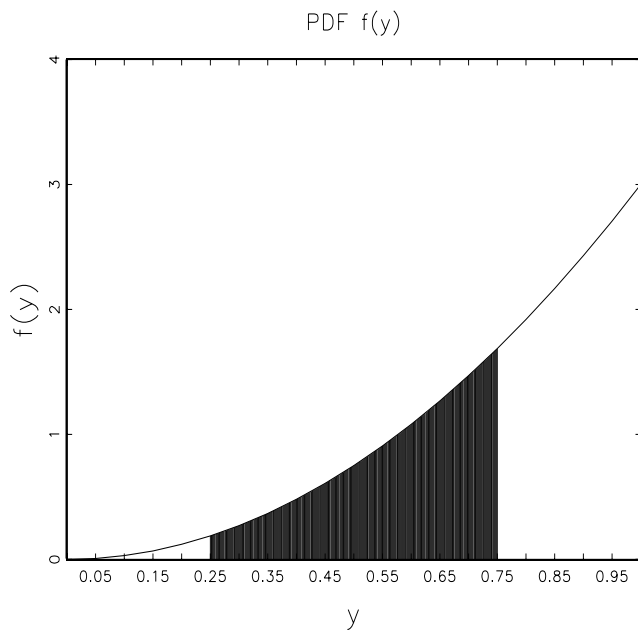


Figure 6: The plot of the pdf $f(y) = 3y^2$. The shaded region corresponds to the probability the random variable Y lies between 0.25 and 0.75.

continuous distribution is the normal distribution which is discussed separately in Section 3.5. We list here some other important continuous distributions.

Uniform Distribution.

The simplest continuous distribution is the uniform distribution whose pdf is a constant function over a finite interval. We say that a random variable Y has a uniform distribution on the interval $[a, b]$ if its pdf is

Uniform pdf: $f(y) = \frac{1}{b-a}$ for $a \leq y \leq b$, and zero otherwise.

The uniform distribution is useful for random variables that are “equally likely” to take any value on a given interval. One important application of the uniform distribution on the interval $[0, 1]$ is in random number generation which is used for Monte Carlo studies. Note that the mean of the uniform distribution is $\mu = (a + b)/2$, the midpoint of the interval. (Note – there is also a discrete uniform distribution which places an equal probability on each value assumed by the random variable.)

Gamma Distributions

Another important class of continuous distributions is the gamma distributions. These are useful for modeling times between events such as in reliability and queuing studies. The pdf for the gamma distribution depends on two parameters, $\lambda > 0$ and $\alpha > 0$. Thus, the gamma distributions represent a *family* of probability distributions because we get a different distribution for different values of λ and α . Also, gamma random variables are positive valued and the pdf is skewed to the right. The pdf for the gamma distribution is

Gamma pdf $f(y) = \frac{\lambda^\alpha}{\Gamma(\alpha)} y^{\alpha-1} e^{-\lambda y}$, for $y > 0$, and zero otherwise.

This pdf uses the gamma function defined by

$$\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} dy.$$

The mean of the gamma distribution is α/λ and the variance is α/λ^2 . The parameter α is known as a *shape* parameter because different values of α change the shape of the pdf. Figure 7 shows a plot of three gamma pdf's; in each case $\lambda = 1$ and values for α were 1, 2 and 3 as indicated in Figure 7.

A special case of the gamma distribution is the *Exponential Distribution* which results when $\alpha = 1$. The pdf for the exponential distribution has a simple form:

Exponential pdf $f(y) = \lambda e^{-\lambda y}$ for $y > 0$, and zero otherwise.

From the mean and variance formulas for the gamma, we deduce that the mean and variance of the exponential distribution is $1/\lambda$ and $1/\lambda^2$ respectively. These quantities can be found directly applying the definition the expectation and variance and integrating by parts.

Example. Lucas (1985) studied times between accidents for a 10 year period at a Dupont facility. Figure 8 shows a histogram of the $n = 178$ times between accidents.

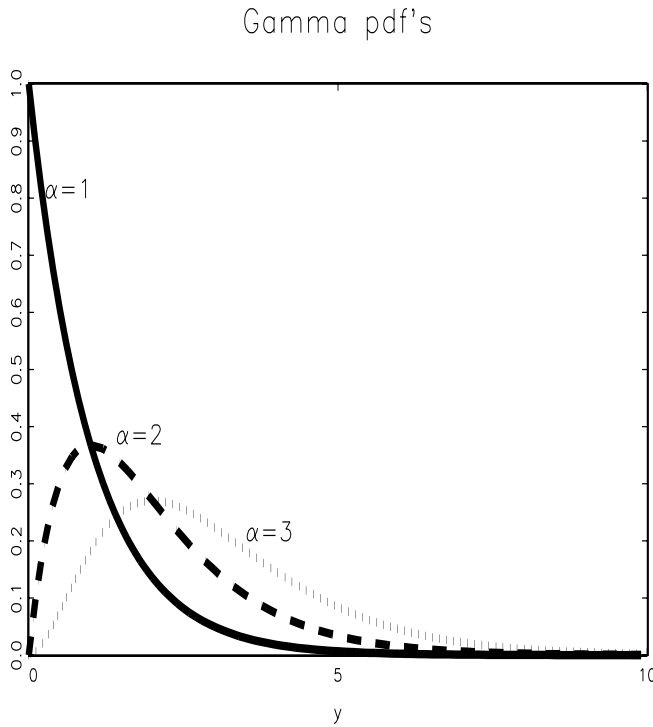


Figure 7: Plots of the gamma pdf for $\lambda = 1$ and $\alpha = 1, 2$ and 3 .

Note that the distribution is very strongly skewed to the right. Often skewness in a distribution occurs when there is a tight boundary and values can only occur to one side or the other of the boundary. In these situations, the distribution then tends to skew in the direction away from the boundary. In Figure 8, times between accidents can only be non-negative and in the data set there are several times near zero.

The time between accident example prompts the following question: how do we determine the appropriate value of λ when modeling the data using an exponential distribution? This is where statistics comes in to play. The mean of the exponential distribution is $1/\lambda$ and it seems reasonable to estimate this mean using the average of the $n = 178$ times from the example. This is exactly what was done to generate Figure 8. However, λ is a population mean and we are estimating that using a sample from the population. Therefore, there is some uncertainty of the true value of λ . A statistical analysis of the data needs to quantify the degree of uncertainty associated with the estimated value. For instance, we would be a lot less certain of the true value of λ if we had estimated it using only $n = 10$ data points say.

3.4 A Brief Note on Transformations.

It is quite common for interest to lie not directly with a measured random variable, but with some transformation of the random variable. A simple transformation of a random variable Y is a linear transformation:

$$X = a + bY,$$

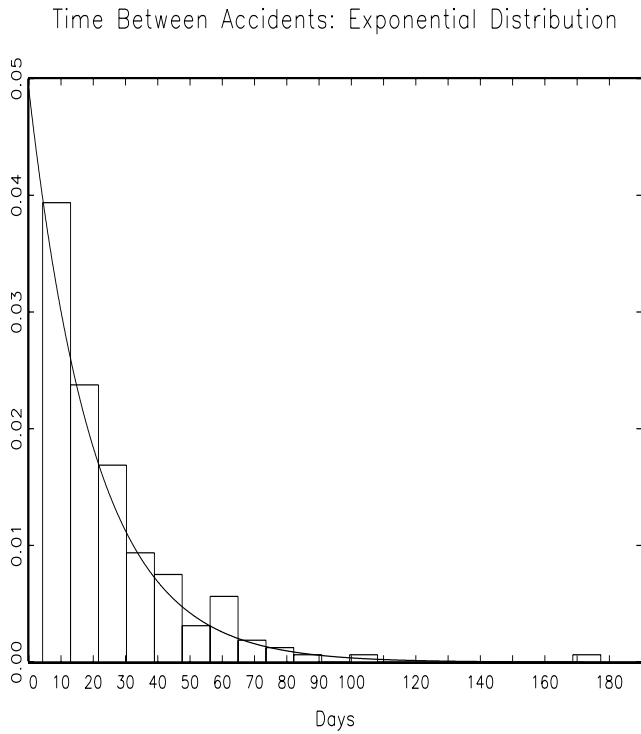


Figure 8: Histogram of times between accidents. Overlaid is a pdf of the exponential distribution with $\lambda = 0.05$.

where $a \neq 0$ and b are constants. For example, Y may represent temperature in Fahrenheit and we can use a linear transformation to change the scale to Celsius. If we want the mean of $X = a + bY$, we simply compute

$$E[X] = E[a + bY] = a + bE[Y].$$

Thus, if μ is the expectation of Y and $X = a + bY$, then $E[X] = a + b\mu$.

The variance of a linear transformation $X = a + bY$ can be computed as follows:

$$\begin{aligned} \text{var}(X) &= \text{var}(a + bY) \\ &= E[\{(a + bY) - (a + b\mu)\}^2] \\ &= E[\{b(Y - \mu)\}^2] \\ &= b^2 E[(Y - \mu)^2] \\ &= b^2 \sigma^2, \end{aligned}$$

where σ^2 is the variance of Y . Therefore,

$$\text{var}(a + bY) = b^2 \text{var}(Y).$$

The important point to note from this formula is that adding a constant a to a random variable simply translates the distribution and does not effect the variability.

However, multiplying a random variable by a constant b changes the variance by a factor of b^2 .

More generally, if $g(y)$ is a function, then $g(Y)$ is also a random variable and one can show that

$$E[g(Y)] = \int_{-\infty}^{\infty} g(y)f(y)dy$$

where $f(y)$ is the pdf of Y .

Example. Suppose that the length of a crack in a fiberglass panel follows a uniform distribution on the interval $[0, 10]$ feet. Let Y denote the size (in feet) of the crack. Then the pdf of Y is

$$f(y) = 1/10, \text{ for } y \in [0, 10],$$

and zero otherwise. Furthermore, the expected value of Y is $\mu = (0 + 10)/2 = 5$. Thus, the average crack length is 5 feet. Also, the variance of Y can be computed as:

$$\begin{aligned} \text{var}(Y) &= E[Y^2] - \mu^2 \\ &= \int_0^{10} (1/10)y^2 dy - 5^2 \\ &= (1/30)(10^3 - 0^3) - 25 \\ &= 25/3. \end{aligned}$$

Now, suppose we define a random variable X to be the crack length in inches. Then

$$X = 12Y$$

and $E[X] = 12E[Y] = 12(5) = 60$ inches, and $\text{var}(X) = 12^2\text{var}(Y) = 144(25/3) = 1200 \text{ in}^2$.

Suppose that the cost of fixing the crack depends on the length of the crack. The cost rises slowly as the length of the crack increases from zero and then the cost rises rapidly as cracks get bigger and bigger. In particular, if the length of the crack is Y , then the cost of fixing the crack is $C = g(Y) = Y^2$ dollars. Now we have a new random variable C representing the cost of fixing the crack. What is the average cost of fixing a crack?

$$\begin{aligned} E[C] &= E[g(Y)] \\ &= \int_0^{10} g(y)f(y)dy \\ &= \int_0^{10} y^2(1/10)dy \\ &= 33.\bar{3} \text{ dollars.} \end{aligned}$$

Its important to note that $g(\mu)$ is not in general equal to $E[g(Y)]$. For instance, in this example, $g(\mu) = g(5) = 5^2 = 25$ is not equal to the expected cost of \$33.33. Therefore, we have in general that

$$g(\mu) \neq E[g(Y)].$$

Equality does hold when $g(y)$ is a linear function.

Problems

1. Let X denote a continuous random variable with pdf $f(x) = x^3/4$ for $0 < x < 2$ and zero otherwise. Find the following:
 - a) $P(X < 0.5)$
 - b) $P(X > 1)$
 - c) $P(X > 4)$
 - d) $E[X]$
 - e) The variance of X .
 - f) $E[X^2 - 2X + 3]$

2. Suppose the random variable Y represents the proportion of alcohol in a solvent. The pdf of Y is $f(y) = k\sqrt{y}$, for $0 < y < 1$ and zero otherwise where k is a constant.
 - a) Find the value of k so that $f(y)$ is a legitimate pdf.
 - b) Find the probability that the proportion of alcohol in the solvent is less than 0.40.
 - c) Find the mean amount of solvent $E[Y]$.
 - d) Find the variance σ^2 of Y .

3. Suppose an engineer working on a road construction project is concerned about managing water runoff after rainfalls. Data on rainfall was collected by the Automated Flood Warning System (AFWS) (<http://www.afws.net/>) giving the amount of rain collected in rain gauges in inches. The following is a set of $n = 55$ rainfall measurements collected at a station in Allen County, Ohio during 1999 arranged in ascending order:

0.02, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04, 0.04,
0.04, 0.04, 0.04, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08, 0.08,
0.08, 0.08, 0.12, 0.12, 0.12, 0.16, 0.16, 0.16, 0.16, 0.16,
0.20, 0.20, 0.20, 0.20, 0.24, 0.28, 0.32, 0.32, 0.32, 0.36,
0.40, 0.40, 0.44, 0.48, 0.52, 0.56, 0.56, 0.56, 0.59, 0.60,
0.72, 1.20, 1.23, 1.59, 2.22.

- a) Make a histogram of this data. Use measurement classes of $0 - 0.1, 0.1 - 0.2, 0.2 - 0.3$, etc. Describe the shape of the rainfall distribution. Re-scale the heights of the histogram rectangles so that the total area under the histogram equals one.
 - b) The average (i.e. the sample mean) of the $n = 55$ rainfalls is $\bar{y} = 0.312$. If we model the rainfall distribution using an exponential distribution, find an estimate of the parameter λ in the exponential pdf by noting the relation between the exponential distribution mean and λ . Sketch the resulting exponential pdf overtop the histogram from part (a) and comment on whether or not it seems to be a reasonable model for the rain data.
 - c) Using the pdf from part (b), compute the probability of a rainfall of $1/2$ inch or more. Compare your answer with the actual proportion of rainfalls from the data set above that exceed 0.50 .
 - d) What is the average amount of rainfall according to this exponential model? That is, compute $E[Y]$.
 - e) Compute the variance σ^2 of Y .
4. The random variable Y represents the amount of time (in years) that a gasket will work before it starts leaking. Suppose the pdf of Y is

$$f(y) = 2e^{-2y}, \quad y > 0.$$

Find the following:

- a) What is the probability the gasket will last more than 1 year?
- b) What is the probability that the gasket will need to be replaced in less than half a year?
- c) What is the average time it takes for these gaskets to start leaking?

3.5 The Normal Distribution.

The most important distribution in probability and statistics is the normal distribution. It is also known as the *Gaussian* distribution in honor of the famous mathematician Karl Frederick Gauss (1777-1855). There are two reasons why the normal distribution is so important.

1. Many random phenomena are modeled well by the normal probability distribution.
2. The most commonly used method to summarize data is to use the average. Averages tend to behave approximately like normal random variables (see Section 3.6).

The normal distribution has a symmetric bell-shaped pdf given by the following formula:

Normal pdf
$$f(y) = \frac{1}{\sqrt{2\pi}\sigma} e^{-(y-\mu)^2/(2\sigma^2)}, \text{ for } -\infty < y < \infty.$$

The normal pdf is centered at μ , the mean (or expected value) of the distribution. Using the definition of variance and integrating, it follows that σ^2 is the variance of the normal distribution.

Notation. If Y is a normal random variable with mean μ and variance σ^2 , then we shall write:

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

Thus, if $Y \sim N(2, 4)$, then Y has a normal distribution with mean equal to 2 and variance equal to 4.

Figure 9 shows three normal pdf's for (i) $\mu = -1, \sigma = 1/4$, (ii) $\mu = 0, \sigma = 1$ and (iii) $\mu = 2, \sigma = 2$. As σ gets smaller, the spread of the distribution decreases and the bell-shaped curve becomes a "steep mountain." On the other hand, as σ increases, the distribution becomes more spread out. Changing μ amounts to simply picking up the pdf and moving it so that it is centered over μ .

There is no closed form formula for the normal cdf. In order to compute normal probabilities, numerical integration is required. However, many software packages have normal cumulative distribution functions built-in. Otherwise, tabled values (see Appendix, page 199) are usually given in the special case of $\mu = 0$ and $\sigma = 1$, the *Standard Normal*.

3.5.1 The Standard Normal.

In order to compute normal probabilities for a normal random variable Y , we typically *standardize* Y first by centering it at zero and rescaling it to have unit variance. This is summarized by the following result:

Fact: If $Y \sim N(\mu, \sigma^2)$, then

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

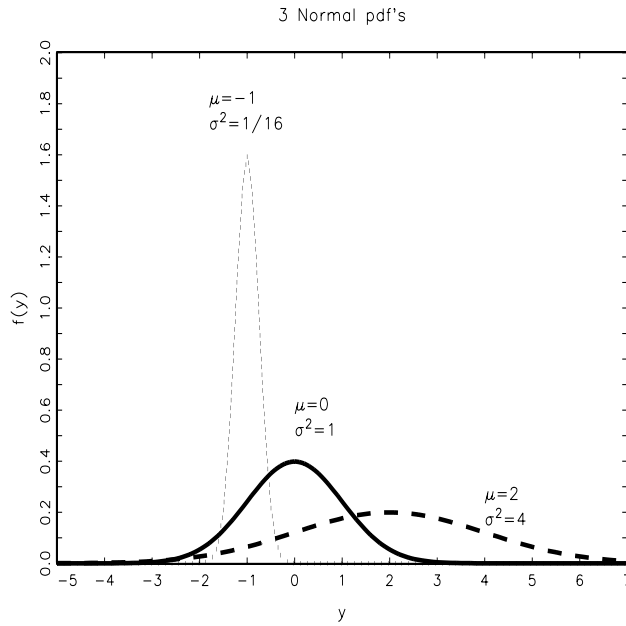


Figure 9: Three normal pdf's for varying values of μ ($-1, 0$ and 2) and varying values of σ^2 ($1/16, 1, 4$).

is a normal random variable with mean zero and variance one. We call the random variable Z the *Standard Normal* random variable (and its distribution is called the standard normal distribution).

Thus, a standard normal probability table can be used to compute any normal probability because if $Y \sim N(\mu, \sigma^2)$, then

$$\begin{aligned} P(a < Y < b) &= P((a - \mu)/\sigma < (Y - \mu)/\sigma < (b - \mu)/\sigma) \\ &= P((a - \mu)/\sigma < Z < (b - \mu)/\sigma) \end{aligned}$$

where Z is a standard normal random variable. Note that the table only gives probabilities to four decimal places of accuracy.

To illustrate the computations of normal probabilities, first we illustrate using a standard normal probability table to compute standard normal probabilities. Cumulative probabilities for the standard normal probability distribution can be found on page 199 in the Appendix.

Example. Find $P(Z < 1.00)$. That is, find the area under the standard normal pdf to the left of $z = 1.00$. Going to the standard normal probability table in the appendix (page 199), we can read the answer directly from the table as

$$P(Z < 1) = 0.8413$$

which can be found in the second column of the table under the row labeled $z = 1.0$. The other columns of the table give probabilities for values of z to the 1/100th decimal

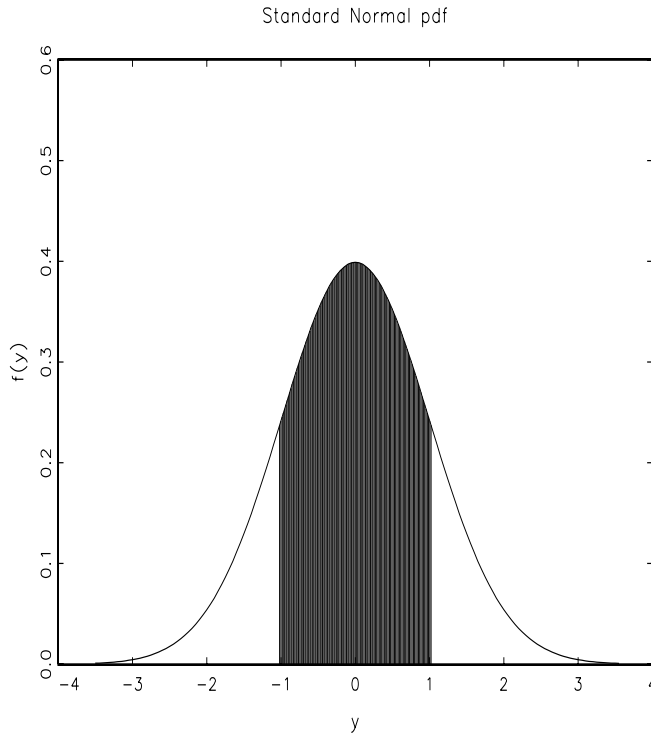


Figure 10: A plot of the standard normal pdf. The shaded region corresponds to $P(-1.02 < Z < 1.02)$.

place of accuracy. For instance,

$$P(Z < 1.02) = 0.8461,$$

which can be found in the $z = 1.0$ row under the 4th column for 0.02.

Example. For another illustration, find $P(-1.02 < Z < 1.02)$. This corresponds to the shaded region under the standard normal pdf shown in Figure 10. This probability can be computed by noting that $P(-1.02 < Z < 1.02) = P(Z < 1.02) - P(Z < -1.02) = 0.8461 - 0.1539 = 0.6922$.

To compute a probability for a general normal, we first standardize and then use the standard normal cdf table. This is illustrated in the next example.

Example. Consider the LTG measurement from the Swiss Army head dimension data. Assume the distribution is normal with mean $\mu = 138.8$ and standard deviation $\sigma = 5.64$. Find the probability that a randomly selected soldier has an LTG measurement exceeding 150 mm. Let Y denote a random variable following the LTG distribution, i.e. $Y \sim N(138.8, 5.64^2)$. The probability $P(Y > 150)$ is indicated by the shaded region in Figure 11. $P(Y > 150)$ is computed as follows:

$$P(Y > 150) = P\left(\frac{Y - \mu}{\sigma} > \frac{150 - \mu}{\sigma}\right) \text{ (standardize)}$$

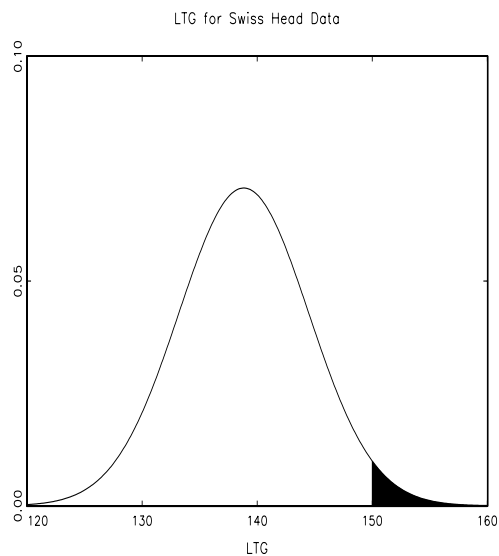


Figure 11: The normal distribution representing the LTG head measurement from the Swiss Army data with mean $\mu = 138.38$ and standard deviation $\sigma = 5.64$. The shaded region under the pdf represents the probability that the LTG measurement exceeds 150 mm.

$$\begin{aligned} &= P\left(Z > \frac{150 - 138.8}{5.64}\right) \\ &= P(Z > 1.9858) \\ &= 1 - P(Z \leq 1.9858) \quad (\text{Law of Complements}) \\ &\approx 1 - 0.9765 \quad (\text{Standard normal table lookup}) \\ &= 0.0235. \end{aligned}$$

Thus, approximately 2% of the soldiers have LTG measurements exceeding 150 mm. It is interesting to note that out of the $n = 200$ observations in this data set, exactly 5 of them had LTG measurements exceeding 150. 5 out of 200 is 0.025 which is extremely close to the probability produced by the normal probability model.

Problems

1. Let Z denote a standard normal random variable. Find the probabilities of the following events. In each case, sketch the standard normal pdf and shade the area under the pdf curve corresponding to the event. Use the cumulative standard normal tables in the Appendix (page 199).
 - a) $P(Z \leq 2.3)$
 - b) $P(Z \leq 2.36)$
 - c) $P(-1.96 < Z < 1.96)$
 - d) $P(|Z| > 2.01)$

2. Find the 90th percentile of the standard normal distribution. That is, find the value z_0 so that $P(Z \leq z_0) = 0.90$.

3. Find the 10th percentile of the standard normal distribution.

4. A quality control study on gears was conducted. One of the variables of interest is the gear diameter. The mean diameter is $\mu = 1$ inch and the standard deviation is $\sigma = 0.006$ inches. Assume the distribution of gear diameters varies according to a normal distribution. Find the following:
 - a) What proportion of gears have diameters exceeding 1.01 inch?
 - b) What proportion of gears have diameters less than 0.99 or greater than 1.01 inches?
 - c) The investigators want to determine an interval about μ containing 95% of the gear diameters. Let Y denote a random variable corresponding to the diameter of a randomly chosen gear ($Y \sim N(1, 0.006^2)$). Determine the number c so that $P(\mu - c < Y < \mu + c) = 0.95$.

5. Measurements were obtained on a sample of $n = 50$ electrodes in a quality control study (Flury 1997). One of the measurements was the width of the head of the electrode. The data from this investigation are below (note – the original data was linearly transformed for reasons of confidentiality).

56	56	56	57	57	58	58	58	58	58
58	58	58	58	59	59	59	59	59	59
59	59	59	59	60	60	60	60	60	60
60	60	60	60	60	60	60	60	61	61
61	61	61	61	62	62	62	63	64	64

- a) What is the median of these 50 measurements?
- b) Construct a histogram of these 50 measurements. Describe the shape of the distribution.

- 6) This is a continuation of problem 5. Assume the mean and standard deviation for the electrode head widths are $\mu = 59.5$ and $\sigma = 1.8$ respectively. Let Y denote the head width of a randomly chosen electrode. Assuming $Y \sim N(59.5, 1.8^2)$, find the following probabilities and sketch the normal pdf and shade the area under the curve corresponding to the probability.
- a) $P(Y \geq 57)$
 - b) $P(58 \leq Y \leq 60)$.
7. A study was conducted at NIST examining the thickness of mica washers. Suppose the mean thickness is $\mu = 0.125$ inches and the standard deviation of thicknesses is $\sigma = 0.005$ inches. Furthermore, suppose the washer thicknesses vary according to a normal distribution.
- a) Find the probability that a randomly selected washer has a thickness exceeding 0.130 inches.
 - b) Find the probability that the thickness of a randomly selected washer differs from the mean thickness by more than 0.01 inches.

3.6 Random Behavior of Means

In the previous problems, it was assumed the μ and σ^2 were known. However, in practice these quantities are unknown and must be estimated from the data. Probability distributions are defined by parameters, such as the λ and α from the gamma distribution. In every case, we need to use the data to obtain estimates for these parameters. Typically we assume our data is a random sample.

Definition. A *random sample* is a collection of independent observations Y_1, Y_2, \dots, Y_n that all have the same distribution. We say that the Y_i 's are i.i.d. meaning independent and identically distributed.

Note that once the data are collected we have a collection of n fixed numbers which we typically denote by using lower-case letters: y_1, y_2, \dots, y_n corresponding to the realized values of our random sample. When hypothesis testing was introduced using the binomial distribution, we identified a test statistic in that setting to be the number of successes. We now give a more general definition of *statistic*.

Definition. A *statistic* is a function of the data that can be computed without knowing the true value of any parameters.

The two statistics encountered most frequently are the sample mean and sample variance.

Definition. Given a random sample y_1, y_2, \dots, y_n , the *sample mean*, denoted \bar{y} , is simply the average of the observations:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

and the *sample variance*, denoted s^2 is

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2.$$

The square-root of s^2 is the *sample standard deviation*.

The sample variance is an average of the squared deviations of each observation from the mean. A natural question to ask at this point is why do we divide by $n-1$ in the definition of s^2 ? We shall give the reason in the next section.

To illustrate the computation, we return to the Swiss head dimension example. To make matters simple, we shall only consider the first 5 of the LTG measurements given here:

$$y_1 = 143.6, \quad y_2 = 143.9, \quad y_3 = 149.3, \quad y_4 = 140.9, \quad y_5 = 133.5.$$

The sample mean then is

$$\bar{y} = \frac{1}{5}(143.6 + 143.9 + 149.3 + 140.9 + 133.5) = 142.24.$$

The sample variance is computed as

$$\begin{aligned} s^2 &= \frac{1}{5-1} \{ (143.6 - 142.24)^2 + (143.9 - 142.24)^2 + \\ &\quad (149.3 - 142.24)^2 + (140.9 - 142.24)^2 + (133.5 - 142.24)^2 \} \\ &= 33.158. \end{aligned}$$

The sample standard deviation is $s = \sqrt{33.158} = 5.76$.

3.6.1 A Note on Robust Measures of Center.

The mean is perhaps the most used measure of “central tendency” since it is the center of gravity of a distribution. However, the mean is not always the best measure of center for data sets, particularly strongly skewed data sets or data sets with outliers. For such data sets, the median is often a better measure of central tendency. For instance, we often see the median reported when data on home prices and salaries are summarized since these types of data sets tend to be skewed to the right and often have extreme observations. To illustrate, suppose the following are $n = 10$ home prices (in thousands of dollars) from a small neighborhood

112, 115, 116, 120, 120, 122, 122, 122, 124, 400.

All but one of the home prices are in roughly the same range with one super expensive home. The median can be determined easily by looking at the middle two observations: median = \$121 thousand dollars which seems to be a reasonable measure of central tendency. However, the mean home price is \$147,300! All but one of the home prices is below average. The average is highly influenced by the single extreme home price of 400 thousand dollars, whereas the median is not pulled away from the center by this expensive home. We say that the median is a robust measure of central tendency. For symmetric populations, the mean and median are equal. However, for skewed distributions the mean is often pulled away from the center in the direction of the skewness while the median tends not to be so strongly affected by skewness. Hence, the median is usually a better measure of central tendency than the mean for skewed distributions.

3.6.2 Sampling Distributions.

There are two fundamentally different types of variability in statistics. We saw in Chapter 1 that there is a natural variability in almost every population. The source of the variability is due to the fact that no two items are exactly alike. We have seen how to define this variability for a population by the mathematical expression for σ^2 and by s^2 for a sample.

We now introduce the other notion of variability which students often find more difficult to grasp. Note that in the Swiss gas mask example, μ the mean value of LTG is unknown and must be estimated. The sample mean from the first five observations

computed above is 142.24. If we use the entire data set of $n = 200$ observations, the sample mean comes out to be $\bar{y} = 138.834$. However, suppose the anthropologist had sampled a different set of 200 soldiers and computed \bar{y} ? They would have arrived at a different value because they were using different measurements. We would expect the values to be close to each other though if the sample size is large. The main point is that \bar{y} varies depending on the sample that was obtained at random. Let the upper-case Y_i 's denote the random variables representing the possible values in our random sample and let the lower-case y_i 's denote the observed values of the random variables once the data has been collected. Then \bar{Y} , the sample mean, can vary depending on the sample that is obtained by random sampling. This line of reasoning illustrates one of the main points:

\bar{Y} is a random variable.

Any time we compute a statistic based on random variables Y_1, Y_2, \dots, Y_n , from a random sample, the result is also a random variable. If we want to use \bar{y} to estimate μ , we need to know how the random variable \bar{Y} behaves. We need to know something about its *sampling distribution*:

Definition. The probability distribution of a statistic is called its *sampling distribution*.

There are a couple of facts we can record right away regarding the behavior of the sample mean. First, \bar{Y} is *unbiased* for μ . What this means is that \bar{Y} does not consistently under nor over-estimate μ . A statistic is said to be an unbiased estimator for a parameter if the expected value of the statistic is equal to the parameter. To say \bar{Y} is unbiased for μ means

$$E[\bar{Y}] = \mu.$$

The proof of this requires some background in jointly considering the distribution of all the Y_i 's together. However, the linearity of the expectation still holds in this case and we can write

$$\begin{aligned} E[\bar{Y}] &= E\left[\frac{1}{n} \sum_{i=1}^n Y_i\right] \\ &= \frac{1}{n} \sum_{i=1}^n E[Y_i] \\ &= \frac{1}{n} \{\mu + \mu + \dots + \mu\} \\ &= \mu. \end{aligned}$$

One can also show that the statistic S^2 is unbiased for σ^2 when we have a random sample. That is,

$$E[S^2] = \sigma^2.$$

This is the reason that S^2 is defined by dividing by $n - 1$ instead of n , so that S^2 will be unbiased. If we divided by n (instead of $n - 1$) then the sample variance would be too small on average.

Another fact about the sample mean that is very important is that the variance of its sampling distribution is equal to the population variance divided by the sample size. Let $\sigma_{\bar{y}}^2$ denote the variance of \bar{Y} . Then

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}.$$

Taking the square root of this gives the *standard error* of the sample mean:

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}. \quad (1)$$

In practice, the standard error, as well as the mean must be estimated because σ , like μ , will not be known. An estimated standard error can be obtained by replacing σ in (1) by the observed sample standard deviation s :

$$\text{Estimated Standard Error of } \bar{y} : \hat{\sigma}_{\bar{y}} = \frac{s}{\sqrt{n}}.$$

This result is somewhat intuitive. Suppose we want to estimate the average height μ of students at a university. If we take repeated samples of size $n = 5$ say and compute the sample mean of each sample, we may expect to see widely varying averages. However, if we take repeated samples of size $n = 100$, then we would not expect to see as much variability in the average heights. In other words, a larger sample provides us more information and therefore the estimated mean will not vary as much in repeated sampling.

We have established that the sample mean is unbiased for the population mean and the standard error of the mean is the population standard deviation divided by \sqrt{n} . The third main point regarding the sampling distribution of \bar{y} is given by the *Central Limit Theorem*.

The Central Limit Theorem. *Let Y_1, Y_2, \dots, Y_n denote a random sample from a distribution with mean μ and variance $\sigma^2 < \infty$. Then for sufficiently large n , the sampling distribution of \bar{Y} is approximately normal.*

The central limit theorem tells us that sample means behave approximately like normal random variables provided we have a large enough sample size. We can then compute probabilities for \bar{Y} by noting that

$$Z = \frac{\bar{Y} - \mu}{\sigma_{\bar{y}}}$$

has approximately a standard normal distribution. Of course, this expression requires that we know μ . However, when we discuss hypothesis testing, we can plug a known value of μ specified by the null hypothesis and then see if z behaves approximately like a standard normal random variable. The above expression also requires that we know σ . We shall see how to resolve this problem in the next section.

Figure 12 illustrates the central limit theorem effect. Suppose the random variable Y represents the amount of rain (measured in inches) and that Y has an exponential

Central Limit Theorem Effect

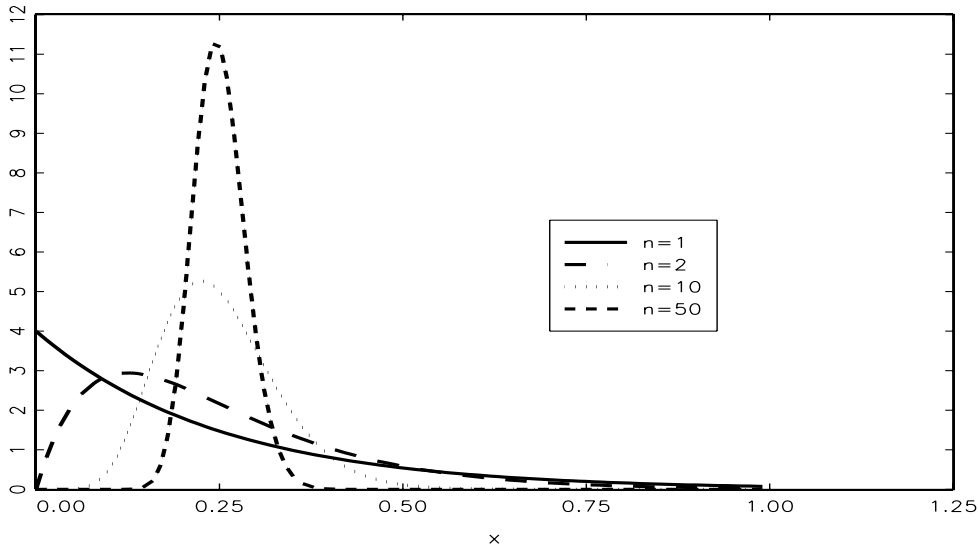


Figure 12: An illustration of the central limit theorem effect. The underlying distribution is exponential. This plot shows the sampling distribution of \bar{Y} for samples of size $n = 1, 2, 10$ and 50 . Note that the sampling distributions of \bar{Y} for each sample size are all centered about $\mu = 1/4$.

distribution with pdf $f(y) = 4e^{-4y}$ for $y > 0$. The average rainfall is $\mu = 1/4$. The distribution of Y is skewed quite strongly to the right as can be seen in Figure 12 for $n = 1$. Suppose everyone in a large class of students random pick two rainfalls and averages them. Then the distribution of these averages based on $n = 2$ rainfalls would have a distribution shown in Figure 12 for $n = 2$. This is the sampling distribution of \bar{Y} for $n = 2$ which is still quite strongly skewed to the right and centered over $\mu = 1/4$. If each student averages $n = 10$ rainfalls selected at random, then the distribution of the \bar{Y} 's for $n = 10$ will look much more symmetric and bell-shaped as shown in Figure 12. The \bar{Y} distribution is still centered over $\mu = 1/4$ and the spread is much less than for $n = 1$ or 2 . For $n = 50$, the sampling distribution of \bar{Y} looks very close to the normal distribution in Figure 12.

The central limit theorem says that \bar{Y} behaves like a normal random variable provided the sample size n is sufficiently large. What does sufficiently large mean? The answer is that it depends on what the underlying distribution looks like. If the underlying distribution is normal, then \bar{Y} will have an exact normal distribution for all sample sizes. If the underlying distribution is approximately normal as evidenced by a bell-shaped histogram with no extreme outliers, then the distribution of \bar{Y} may be well-approximated by a normal distribution for relatively small sample sizes ($n = 5, 10$, etc.). However, if the underlying distribution is strongly skewed, or has “heavy” tails, then a larger sample size may be required to guarantee a good normal approximation. A rule of thumb often seen in textbooks is that if $n \geq 30$, then \bar{Y} will be approximately normal. However, it is not difficult to find examples where this rule of thumb fails.

It is always a good idea to plot your data and see if it looks approximately normal or not.

Example. In the cup-a-soup example, suppose the average fill-weight is set to be 238 oz. per cup on average. From the sample of $n = 969$ cups, the observed sample mean is $\bar{y} = 238.401$. The sample standard deviation is $s = 0.9681$ oz. If the true mean fill-weight μ for this production process is indeed 238 oz., how likely is it that we would observe a sample mean of 238.401 oz. or more? Let us assume for the moment that the true σ is equal to the sample standard deviation $s = 0.9681$. Thus, we want to compute $P(\bar{Y} \geq 238.401)$ assuming $\mu = 238$. The observed sample mean looks to be very close to the hypothetical value of 238. Because the sample size is so large, \bar{Y} will have an approximate normal distribution by the central limit theorem, even though the distribution of soup fill-weights is skewed to the left (see Figure 3). Let z denote a standard normal random variable as before. Computing, we get

$$\begin{aligned} P(\bar{Y} \geq 238.401) &= P\left(\frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \geq \frac{238.401 - \mu}{\sigma/\sqrt{n}}\right) \\ &\approx P\left(Z \geq \frac{238.401 - \mu}{\sigma/\sqrt{n}}\right) \text{ (by the central limit theorem)} \\ &= P\left(Z \geq \frac{238.401 - 238.000}{0.9681/\sqrt{969}}\right) \text{ (assuming } \mu = 238) \\ &= P(Z \geq 12.8615) \\ &\approx 0. \end{aligned}$$

Therefore, assuming the true mean is 238, we observed a sample mean that is almost 13 standard deviations beyond its mean; or, after standardizing the observed value of \bar{y} , we get a value of 12.8615 which should look like a number produced by the standard normal distribution. However, it is quite rare that the standard normal distribution would produce a number beyond ± 3 and extremely unlikely it would produce a number as big or bigger than 12.8615. This indicates that something went wrong. The computation was predicated on the assumption that $\mu = 238$. Thus, it appears this assumption is false and that the true mean differs from 238.

Problems

1. Air tanks are manufactured for use in welding. The proportion of oxygen in the tanks varies according to a distribution with pdf

$$f(y) = \begin{cases} 3y^2, & 0 < y < 1 \\ 0 & \text{otherwise} \end{cases}.$$

- a) Find the mean oxygen level μ in the tanks.
- b) Find the standard deviation σ for the distribution of oxygen level in the tanks.

2. This problem is a continuation of problem 1. The company is concerned that the proportion of oxygen in the tanks is too low so they conduct a study by measuring proportion of oxygen in $n = 50$ tanks obtaining measurements y_1, \dots, y_{50} .
- Let \bar{Y} denote the sample mean before the data is collected. What is $E[\bar{Y}]$ assuming the distribution of oxygen in the tanks follows the distribution specified in problem 1?
 - What is the standard error $\sigma_{\bar{y}}$ of the sample mean in part (a)?
 - Suppose the observed sample mean comes out to be $\bar{y} = 0.67$. Using the central limit theorem, how likely is it to see sample mean taking the value 0.67 or less if the true mean is equal to the answer you found in part (a) of problem 1?
3. (This problem appeared earlier in the chapter, without part (d) below). The random variable Y represents the amount of time (in years) that a gasket will work before it starts leaking. Suppose the pdf of Y is

$$f(y) = 2e^{-2y}, \quad y > 0.$$

Find the following:

- What is the probability the gasket will last more than 1 year?
 - What is the probability that the gasket will need to be replaced in less than half a year?
 - What is the average time it takes for these gaskets to start leaking?
 - Suppose there are $n = 100$ gaskets in operation at a factory. The amount of time until a leak developed was noted for each of these gaskets. The average time until a gasket leaked for the 100 gaskets was reported as $\bar{y} = 2.12$ years. How likely is it that the sample mean takes a value of 2.12 or greater if Y has the distribution specified above? Justify your answer.
4. A study was conducted at NIST examining the thickness of mica washers. Suppose the mean thickness is $\mu = 0.125$ inches and the standard deviation of thicknesses is $\sigma = 0.005$ inches. Furthermore, suppose the washer thicknesses vary according to a normal distribution.
- Find the probability that a randomly selected washer has a thickness exceeding 0.132 inches.
 - In this study, a sample of $n = 10$ washers were sampled and their thickness recorded. How likely is it that the sample mean \bar{Y} takes a value exceeding 0.132 inches?

3.7 The t -Distribution.

In the previous example, we assumed the variance was known when in fact we estimated it using the sample variance. In practice, when we are trying to estimate or infer something about the true (unknown) value of the population mean μ , it is almost always the case that σ will be unknown as well.

If our random sample Y_1, Y_2, \dots, Y_n comes from a normal distribution $N(\mu, \sigma^2)$, then

$$Z = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} \quad (2)$$

has a normal distribution with mean μ and variance $\sigma_y^2 = \sigma^2/n$. If σ^2 is unknown and we replace it by the sample variance S^2 in the above expression, we get

$$t = \frac{\bar{Y} - \mu}{S/\sqrt{n}}. \quad (3)$$

Replacing a constant σ in (2) by a random variable S to get (3) introduces additional variability and changes the distribution. The statistic t in (3) is no longer standard normal but instead follows what is known as the t -distribution on $n - 1$ degrees of freedom. The t distribution is actually a family of distributions that depends on the sample size n by way of its degrees of freedom (df): $\text{df} = (\text{sample size}) - 1$. The pdf for the t distribution on ν degrees of freedom (when dealing with the sample mean $\nu = n - 1$):

$$f(y) = \frac{\Gamma((\nu + 1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} (1 + y^2/\nu)^{-(\nu+1)/2}, \quad \text{for } -\infty < y < \infty.$$

The t pdf has a shape very similar to the standard normal distribution (symmetric about zero and bell-shaped), but it has heavier tails. Note that the standard normal density dies off very rapidly due to the exponential term $e^{-0.5z^2}$. The t pdf does not die off as rapidly.

As the degrees of freedom grows large ($\nu \rightarrow \infty$), the t distribution becomes more and more like a standard normal distribution. Figure 13 shows a plot of the standard normal pdf along with the t distribution pdf for degrees of freedom equal to 2 and 10.

The definition of the t distribution given above assumes that the random sample was obtained from a normal population. However, the resulting t statistic is quite robust to departures from normality. That is, if the normality assumption does not hold, the t testing procedure will still yield approximately valid results, provided the departure from normality is not too severe. If the underlying population is non-normal but fairly mound-shaped without heavy tails, then the t statistic will follow a t distribution approximately. Always plot the data to assess if the normality assumption is valid or at least approximately valid. When there are only a few observations (i.e. n is small), it is difficult to discern the shape of the distribution. Nonetheless, with even a few data points, one can often determine if there is a problem with strong skewness or outliers. For large sample sizes, we can sustain stronger violations against the normality assumption due to the central limit theorem effect. Note that there are

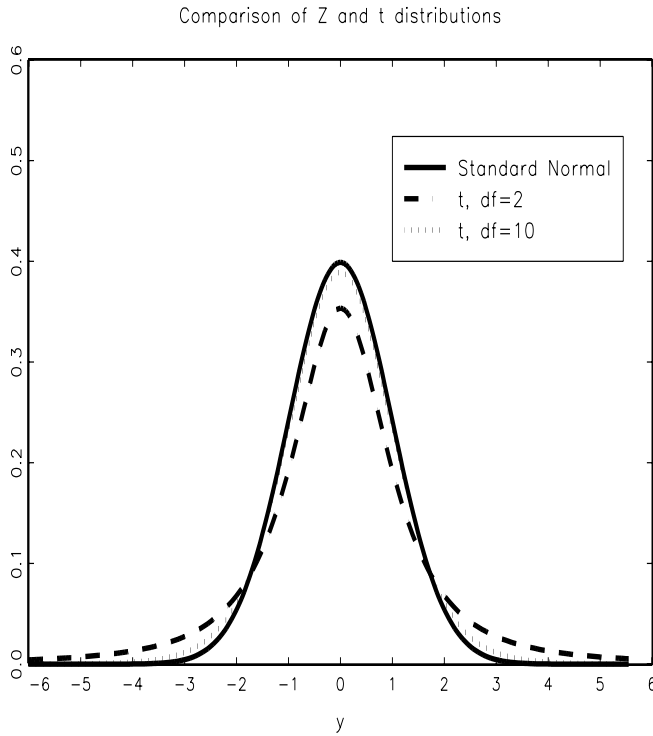


Figure 13: Pdf's of the standard normal and t distributions on 2 and 10 degrees of freedom.

other methods of accessing whether or not the underlying population is normal or not (e.g. goodness-of-fit tests such as the Shapiro-Wilks, and normal probability or Q-Q plots).

3.7.1 t -Critical Values.

In order to use the t distribution for hypothesis testing and estimating population means, it is helpful to define critical values. Given a small probability value $\alpha > 0$, we denote by t_α the value of a t random variable so that the area under the t -density curve to the right of t_α is equal to α . The number t_α is known as a t -critical value. A table of t -critical values can be found in the Appendix on page 201. By definition of the critical value we have

$$P(T > t_\alpha) = \alpha,$$

where

$$T = \frac{\bar{Y} - \mu}{S/\sqrt{n}},$$

assuming our sample comes from a normal population. Figure 14 shows a t pdf on 10 degrees of freedom. Using $\alpha = 0.05$, the $t_{0.05}$ critical value can be found from the t table to be $t_{0.05} = 1.8125$. By definition of critical value, the area under the t pdf to the right of 1.8125 is $\alpha = 0.05$.

The t critical values can be read off the t -table for a given degrees of freedom.

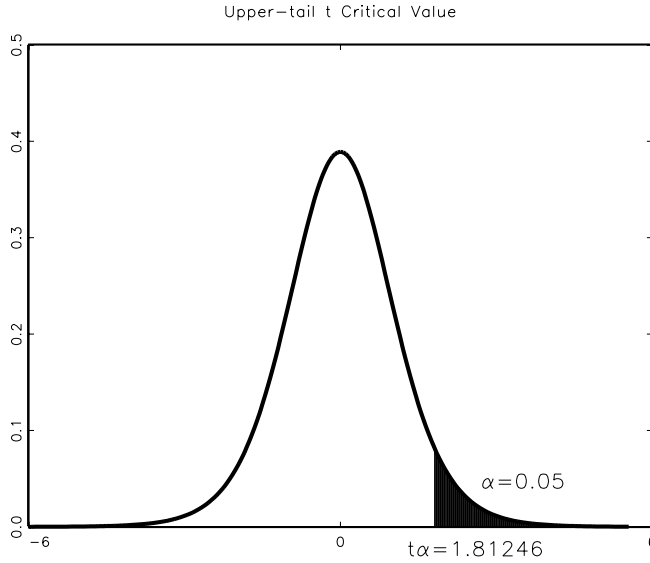


Figure 14: t -distribution critical value for $\alpha = 0.05$ with $\nu = 10$ degrees of freedom is $t_{0.05} = 1.8125$. The area under the t pdf to the right of 1.8125 is $\alpha = 0.05$.

3.8 Hypothesis Testing for the Mean and the t -Distribution.

We return now to hypothesis testing and consider tests concerning the mean of a distribution. The t distribution is typically used to make inferential statements about the mean of a distribution. The null hypothesis states that the mean of the population μ is equal to some hypothesized value μ_0 say:

$$H_0 : \mu = \mu_0.$$

Depending on the application at hand, the alternative hypothesis takes one of the following forms:

$$\begin{array}{ll} H_a : \mu < \mu_0 & \text{One-tailed test} \\ H_a : \mu > \mu_0 & \text{One-tailed test} \\ H_a : \mu \neq \mu_0 & \text{Two-tailed test} \end{array}$$

The test statistic in each case is simply a measure of the standardized difference between the sample mean \bar{y} and the hypothesized mean μ_0 given by the t -test statistic:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} \quad \text{(Test Statistic)} \quad (4)$$

which follows a t distribution on $n - 1$ degrees of freedom when the null hypothesis is true. If we are testing the hypothesis using a significance level α , then the following table summarizes the procedure:

Alternative Hypothesis	Decision
$H_a : \mu < \mu_0$	Reject H_0 if $t < -t_\alpha$
$H_a : \mu > \mu_0$	Reject H_0 if $t > t_\alpha$
$H_a : \mu \neq \mu_0$	Reject H_0 if $t > t_{\alpha/2}$ or $t < -t_{\alpha/2}$

The t -critical values in the above table depend on the degrees of freedom which equals $n - 1$.

Instead of testing the hypothesis at a fixed significance level α one may opt instead to report the p -value of the test. Recall that the p -value is the probability of observing a result (i.e. sample mean) as extreme or more extreme away from the null hypothesis when the null hypothesis is true. p -values near zero are evidence against the null hypothesis and larger p -values indicate that the evidence against the null hypothesis is weak. The following table provides the details on computing p -values for the tests. In this table, T represents a t random variable on $n - 1$ **degrees of freedom**.

Alternative Hypothesis	p -value
$H_a : \mu < \mu_0$	$P(T < t)$
$H_a : \mu > \mu_0$	$P(T > t)$
$H_a : \mu \neq \mu_0$	$2P(T > t)$

The p -value probabilities in the table can be computed by statistical software packages or approximated by using the t -table (see the next example for an illustration).

Finally, the testing procedure based on the t -test statistic assumes that the data are from a normal population. If there are strong violations to the normality assumption, then other testing procedures should be considered, such as *nonparametric tests*. Another approach is to consider a transformation. Often times the original measurement scale at which the data is collected results in a strongly skewed distribution. A common way of dealing with such data is to work with the log-transformed data instead. The following example illustrates the idea.

Example. Albin (1990) studied aluminum contamination in recycled PET plastic from a pilot plant operation at Rutgers University. She collected $n = 26$ samples and measured, in parts per million (ppm), the amount of aluminum contamination. The maximum acceptable level of aluminum contamination, on the average, is 220 ppm. Because the distribution of aluminum amounts is strongly skewed to the right (see Figure 15), the (natural) logarithm transformation of the data is analyzed instead. The right panel in Figure 15 shows a histogram of the log-transformed data. This figure indicates that the log-transformed data appear to follow an approximate normal distribution due to the symmetric bell-shaped pattern. Therefore, the use of the t -testing procedure for the log-transformed data seems reasonable. The goal is to test if the mean $\ln(\text{aluminum contamination})$ is below the maximum acceptable level of $\ln(220) = 5.3936$. The $n = 26$ measurements (before taking logarithms) are

291 222 125 79 145 119 244 118 182
 63 30 140 101 102 87 183 60 191
 119 511 120 172 70 30 90 115

The null hypothesis is that the mean of the log-aluminum contamination μ is at the maximum acceptable level and the alternative hypothesis is that μ is below the maximum acceptable level:

$$H_0 : \mu = 5.3936 \text{ versus}$$

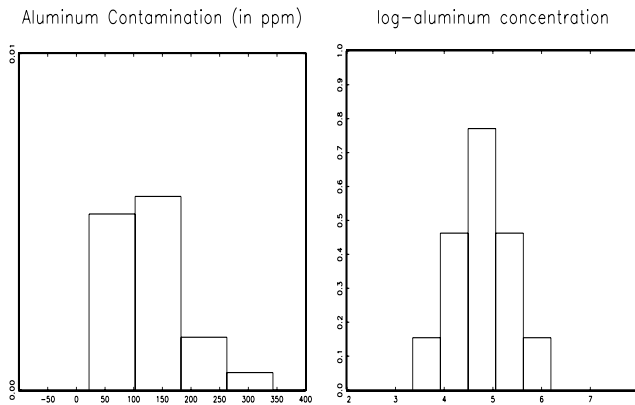


Figure 15: Left Panel shows a histogram of the aluminum contamination measurements which is strongly skewed to the right. The right panel shows a histogram of the log-transformed data showing a symmetric bell-shaped distribution.

$$H_a : \mu < 5.3936.$$

Let us test this hypothesis using a significance level $\alpha = 0.05$. Because there are $n = 26$ observations, the critical region will be based on a t -distribution critical value based on $n - 1 = 26 - 1 = 25$ degrees of freedom. From the above table, we will reject the null hypothesis for this one-tailed test if the test statistic (4) is less than $-t_\alpha = -1.70814$. To compute the test statistic (4), we use with $\mu_0 = \ln(220)$, the null-hypothesis value of the mean. The sample mean (of the log-transformed observations) is $\bar{y} = 4.7729$ and the sample standard deviation is $s = 0.63144$. Plugging these values into the test statistic gives

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{4.7729 - \ln(220)}{0.63144/\sqrt{26}} = -5.0128.$$

Therefore, the observed value of \bar{y} lies about five standard deviations below the hypothesized mean of $\ln(220)$. Because the test statistic t lies in the critical region (i.e. $t = -5.0128 < -t_\alpha = -1.70814$), we reject the null hypothesis at $\alpha = 0.05$ and conclude that the mean log-transformed aluminum concentration is below the acceptable level of $\ln(220)$.

How likely is it for the sample mean \bar{y} to take a value this extreme or more extreme in a direction away from the null hypothesis if the null hypothesis is true? The answer

to this question is given by the p -value:

$$p\text{-value} = P(T \leq -5.0128).$$

Software packages are typically used to find p -values. Using the R-software one can type in the command:

```
pt(-5.0128, 25)
```

to get the p -value. The function “pt” gives cumulative probabilities for the t -distribution. The 25 tells R to use 25 degrees of freedom. The exact p -value found by typing this into R is

$$P(T \leq -5.0128) = 0.00001803$$

which is very small. In other words, if the true mean were $\mu = 5.3936$, then the chance of observing a sample mean value of $\bar{y} = 4.7729$ or less is extremely unlikely (because the p -value is very close to zero). Thus, we have strong evidence for rejecting the null hypothesis and concluding that the mean log-aluminum concentration is below the maximum acceptable level.

If software is not available, an approximate p -value can be found by using the t -table in the appendix. Because the t distribution is symmetric about zero, $P(T < -5.0128) = P(T > 5.0128)$. If we look at the row for df equal to $n-1 = 26-1 = 25$ and scan across, we see the largest recorded value is 3.450 corresponding to an upper-tail probability of 0.001. Since 5.0128 is larger than 3.450, we conclude using the table alone that $p < 0.001$.

3.9 Confidence Intervals.

Many statistical inference problems deal with testing hypotheses. In many other cases, the goal is simply to obtain an estimate of one or more parameters of interest. Returning to the Swiss head dimension example, in the problem of designing gas masks, it would be of interest to know the mean LTG measurement (call it μ) so that appropriate sizes can be determined. From the sample of $n = 200$ samples, we computed the sample mean of $\bar{y} = 138.834$. However, as we noted earlier, the observed value of \bar{y} depends on the random sample of $n = 200$ soldiers that was obtained and that a different sample would have yielded a different value of \bar{y} . Therefore, in the problem of estimating the mean value of *LTG* we not only need a *point estimate* \bar{y} but, in addition, some indication of the variability of the sample mean. It is customary then to report the sample mean as well as its estimated standard error. In the Swiss head data example, $s_{\bar{y}} = s/\sqrt{n} = 5.6424/\sqrt{200} = 0.3990$ mm.

A common approach to parameter estimation is to provide an interval of plausible values for the parameter of interest. These intervals are called *confidence intervals*. The logic behind confidence intervals is given by the following probability illustration. Suppose Y_1, Y_2, \dots, Y_n denotes a random sample from a $N(\mu, \sigma^2)$ population and we want to estimate μ , the unknown mean. We know that

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

follows a t -distribution on $n - 1$ degrees of freedom. For a small probability value of $\alpha > 0$, split α equally in two for the left and right tails of the t -distribution. Then we can write

$$P(-t_{\alpha/2} < T < t_{\alpha/2}) = P(-t_{\alpha/2} < \frac{\bar{Y} - \mu}{S/\sqrt{n}} < t_{\alpha/2}) = 1 - \alpha.$$

Solving for μ inside the inequality in the above probability statement gives

$$P(\bar{Y} - t_{\alpha/2}S/\sqrt{n} < \mu < \bar{Y} + t_{\alpha/2}S/\sqrt{n}) = 1 - \alpha.$$

This expression gives a random interval $\bar{Y} \pm t_{\alpha/2}S/\sqrt{n}$ that includes the true value of μ with probability $1 - \alpha$. Commonly used values for α include $\alpha = .05$ or 0.01 . Once the data is collected, we obtain n realizations of the random variables Y_i that, as we have noted, are denoted by the lower-case y_i 's, $i = 1, 2, \dots, n$. The $n = 200$ numbers in the Swiss head dimension data set are the realizations y_1, y_2, \dots, y_{200} . These are now fixed numbers and it does not make sense to think of them as random variables once the data is collected. Replacing \bar{Y} and S by their realized values gives a $(1 - \alpha)100\%$ *Confidence Interval* for μ :

$$\text{Confidence Interval for } \mu: \quad \bar{y} \pm t_{\alpha/2}(s/\sqrt{n}) \quad (5)$$

where $t_{\alpha/2}$ is the $\alpha/2$ critical value of the t distribution on $n - 1$ degrees of freedom. The *confidence level* is $1 - \alpha$. We can obtain a 95% confidence interval by using $\alpha = 0.05$.

It is important to understand the correct interpretation of a confidence interval. Because we replaced the random quantities \bar{Y} and S by their observed values \bar{y} and s , the confidence interval is not random. It is computed with fixed numbers. Therefore, it does not make sense to say that there is a 95% probability that μ lies in a 95% confidence interval. Because the interval is fixed and μ is a fixed (unknown) parameter, either μ lies in the interval or it does not with probability 0 or 1. How then do we interpret the confidence interval? There is a 95% probability that the random interval will contain μ . If we repeat the experiment over and over again, roughly 95% of the computed intervals will contain μ . Thus, using $\alpha = 0.05$ say, the confidence interval procedure is a procedure that works 95% of the time, i.e. in repeated experiments, 95% of the resulting confidence intervals will contain μ . To illustrate the technique, we now return to the Swiss head example.

Example. For the LTG measurement on the Swiss soldiers, let us find a 95% confidence interval for the mean LTG length μ . The sample mean from the $n = 200$ measurements is $\bar{y} = 138.834$ and $s = 5.624$. If the confidence level is $.95 = (1 - \alpha)$, then $\alpha = 0.05$ and $\alpha/2 = 0.025$. With $n = 200$, we need to obtain the critical value from the t distribution on $n - 1 = 199$ degrees of freedom. The t -distribution table gives the 0.025 critical value for 200 degrees of freedom of $t_{0.025} = 1.9719$ which we shall use since it is very close to the value for degrees of freedom 199. From (5), we compute

$$\bar{y} \pm t_{\alpha/2}(s/\sqrt{n}) = 138.834 \pm 1.9719 \cdot (5.624/\sqrt{200}) = 138.834 \pm 0.7842.$$

This gives an interval of (138.0498, 139.6182). With 95% confidence we estimate that the mean LTG length for male Swiss soldiers lies between 138.0472 and 138.8340 mm.

Notes.

- We can compute confidence intervals for differing levels of confidence. For example, using $\alpha = 0.01$ we would obtain a 99% confidence interval indicating that we are more confident that the interval contains μ than say a 95% confidence interval. However, to be more confident requires a wider interval. Analogously, if I want to be more confident of catching a fish using a net, I need a bigger net. For 99% confidence we have $\alpha = 0.01$ and $\alpha/2 = 0.005$. The corresponding t critical value from the t table is $t_{0.005} = 2.60063$. Using this critical value in (5) results in a confidence interval that is wider than the 95% confidence interval we found in the previous example. Thus, the tradeoff of having more confidence (99% versus 95%) is a less precise interval estimate (i.e. a wider confidence interval).
- In the Swiss head data example, there were several different measurements taken on the head indicating that interest lies in several mean parameters (for the different head measurements). One approach to estimating several means is to compute confidence intervals for the mean of each measurement. However, the widths of the individual intervals need to be lengthened to guarantee the appropriate confidence level (one such method of adjustment is called the *Bonferroni* method). However, a more efficient method is to determine a joint *confidence region*. For two normally distributed measurements (bivariate data), the confidence region is an ellipse; for more than two measurements one obtains a confidence ellipsoid (see Chapter 4 on Multivariate Statistics).
- The idea of a confidence interval extends to other parameters as well. For instance, in a binomial application, we may be interested in estimating the parameter p corresponding to the probability of success. We can often approximate the binomial distribution using a normal distribution if the number of trials n is relatively large and the value of p is not too close in value to 0 or 1 – details are given in Section 3.10.

Problems

1. A standard generator battery lasts for $\mu = 4$ hours on average with standard deviation $\sigma = .3$ hours. An additive is added to the battery in hopes of extending the mean amount of time the battery lasts. $n = 5$ batteries are tested using the new additive. The goal of the study is to determine if the battery additive increases the life of the battery on average.
 - a) State the null and alternative hypotheses for this problem in terms of a model parameter. Define the parameter.

- b) In the context of this problem, what is a type I error?
- c) In the context of this problem, what is a type II error?
- d) Suppose you are asked to perform a statistical test using a significance level $\alpha = 0.05$? What does this mean?
- e) Using $\alpha = 0.05$, what is the critical region for this test? Sketch the t -density and shade the critical region.
- f) What assumptions are necessary for the test to be (approximately) valid?
- g) Suppose the $n = 5$ battery lifetimes using the additive are:

4.1, 4.8, 5.2, 3.9, 4.4.

What are the sample mean and standard deviation of these five measurements?

- h) Compute the t -test statistic to test the hypotheses from part (a). What do you conclude from this test?
- i) Suppose the data was actually given by the following 5 battery lifetimes:

4.37, 4.68, 4.58, 4.49, 4.28.

Compute \bar{y} , s and the t -test statistic for these data. Note that the sample mean comes out exactly as it did for the data in part (e). Why then does the result of the t -test come out differently? How can you explain how two data sets with the same average value lead to two different conclusions?

- j) Compute the p -values for the tests using data from parts (e) and (g). Which p -value is smaller?

2. A study was conducted at NIST where the diameter of gears were measured. The data from one of the batches is in the following list:

1.006, 0.996, 0.998, 1.000, 0.992, 0.993, 1.002, 0.999, 0.994, 1.000

- a) Find the sample mean \bar{y} of these $n = 10$ measurements.
- b) Find the sample standard deviation of these measurements.
- c) The mean diameter μ is supposed to be 1 inch. Perform a t -test to test if the mean diameter differs from $\mu_0 = 1$ using a level of significance $\alpha = 0.05$. Be sure to state the null and alternative hypotheses, compute the t -test statistic, determine the critical region, and state your decision and conclusion.
- d) Find a 99% confidence interval for the mean gear diameter μ using the data above.

3. A study at Sematech was conducted on semiconductor wafers (Hurwitz et al, Section 1.3.1.1). A sample of $n = 16$ measurements were obtained on the thickness of wafers yielding the following data:

2987, 3238, 3459, 3361, 3487, 2884, 3405, 3034, 3281, 3258, 3039, 3191, 3086, 3300, 2980, 3210.

Note that the sample standard deviation for these $n = 16$ measurements is $s = 182.4$.

- a) Find the sample mean of these measurements.
 - b) Find a 95% confidence interval for the mean wafer thickness μ .
 - c) Find a 90% confidence interval for the mean wafer thickness using the same data.
4. Another study conducted at NIST examined the thickness of mica washers. Suppose the mean thickness is supposed to be $\mu = 0.125$. The data from the NIST study is given below:

.123, .124, .126, .129, .120, .132, .123, .126, .129, .128.

- a) Find the sample mean \bar{y} of these $n = 10$ observations.
- b) Find the sample standard deviation s from these measurements.
- c) Perform a hypothesis test to determine if the mean thickness differs from $\mu_0 = 0.125$. using a significance level $\alpha = 0.10$. Be sure to state H_0 and H_a , determine the critical region, compute the t -test statistic and make your decision. In plain English, write out the conclusion of the test.
- d) Find a 99% confidence interval for the mean washer thickness.

3.10 Normal Approximation to the Binomial.

Previously we introduced the discrete binomial distribution. Recall that a random variable Y has a binomial distribution if it represents the number of successes on n independent and identical trials. If we define the following zero-one *Bernoulli* random variables

$$\begin{aligned} Y_1 &= 1 \text{ if trial 1 is a success and zero otherwise} \\ Y_2 &= 1 \text{ if trial 2 is a success and zero otherwise} \\ &\vdots \\ Y_n &= 1 \text{ if trial } n \text{ is a success and zero otherwise,} \end{aligned}$$

then

$$Y = \sum_{i=1}^n Y_i.$$

Dividing by n gives an estimate of the probability p of success: the sample proportion \hat{p} (“ p -hat”) is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^n Y_i.$$

By the central limit theorem, \hat{p} will follow a normal distribution approximately provided n is sufficiently large because \hat{p} is a sample mean. Consequently, the binomial random variable $Y = n\hat{p}$ will also have an approximate normal distribution for large n . The mean and standard deviation of a binomial random variable is $\mu = np$ and $\sigma = \sqrt{npq}$ respectively. Since Y is approximately normal for large n , we have that for $a = 0, 1, \dots, n$

$$P(Y \leq a) = P\left(\frac{Y - np}{\sqrt{npq}} \leq \frac{a - np}{\sqrt{npq}}\right) \approx P\left(Z \leq \frac{a - np}{\sqrt{npq}}\right),$$

where Z is a standard normal random variable. Because Y is discrete and Z is continuous, we can make the approximation a little better by using a *continuity correction factor* by adding 0.5 to a :

$$P(Y \leq a) = P(Y < a + 0.5) \approx P\left(Z < \frac{a + 0.5 - np}{\sqrt{npq}}\right),$$

To illustrate the normal approximation to the binomial, reconsider the cup-a-soup hypothesis test considered earlier where $n = 100$ cups were sampled producing $Y = 90$ successes (where success occurs if the fill-weight in the cup lies between 237 – 239 oz.) In that example, we computed $P(Y \geq 90) = 0.00569638$ assuming $p = 0.80$ using the exact binomial probability computation. In this example, $\mu = np = 100(.80) = 80$, and $\sigma = \sqrt{npq} = \sqrt{100(0.80)(0.20)} = 4$. Using the normal approximation (without the continuity correction), we find

$$P(Y \geq 90) = P\left(\frac{Y - \mu}{\sigma} \geq \frac{90 - \mu}{\sigma}\right)$$

$$\begin{aligned}
&\approx P\left(Z \geq \frac{90 - 80}{4}\right) \\
&= P(Z \geq 2.5) \\
&= 0.00621
\end{aligned}$$

which is quite close to the exact value of 0.00569638.

We can use the normal approximation to the binomial compute other probabilities as well:

$$\begin{aligned}
P(Y = a) &= P(Y \leq a) - P(Y \leq a - 1) \\
&= P(Y < a + 0.5) - P(Y < a - 1 + 0.5) \\
&= P\left(Z < \frac{a + 0.5 - np}{\sqrt{npq}}\right) - P\left(Z < \frac{a - 0.5 - np}{\sqrt{npq}}\right).
\end{aligned}$$

Similarly,

$$P(Y > a) = 1 - P(Y \leq a)$$

and $P(Y \leq a)$ can be approximated as above.

For another illustration, suppose $n = 25$ and $p = 1/2$. Then the exact probability is

$$P(Y = 10) = \binom{25}{10} (1/2)^{10} (1 - 1/2)^{25-10} = 0.09742.$$

The normal approximation gives

$$P\left(Z < \frac{10.5 - np}{\sqrt{npq}}\right) - P\left(Z < \frac{9.5 - np}{\sqrt{npq}}\right) = P(Z < -0.8) - P(Z < -1.2) = 0.09679,$$

which is very close to the exact value.

A question we have not considered yet is how big should n be so that the normal distribution provides a good approximation to the binomial distribution? The previous rule of thumb of $n \geq 30$ does not apply here. If p is close to 0 (or 1), then the binomial distribution will be skewed to the right (or left). Thus, the required number of trials for a good normal approximation depends on the value of p . The usual rule of thumb for the binomial distribution is that n will be sufficiently large for a good normal approximation provided that

$$\begin{aligned}
np &\geq 5 \text{ and} \\
nq &\geq 5.
\end{aligned}$$

Confidence Interval for p . The normal approximation to the binomial can be used to obtain a confidence interval for the success probability p . Because \hat{p} has an approximate normal distribution for large n with estimated standard deviation

$\sqrt{\hat{p}\hat{q}/n}$, a $(1 - \alpha)100\%$ approximate confidence interval for p can be obtained using the same logic as in the case for confidence intervals for μ . The formula is

$$\text{Approximate Confidence Interval for } p: \quad \hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n}$$

Cup-a-Soup example once again... Suppose we want to test a hypothesis about the proportion of cups p that are properly filled in the cup-a-soup example. In particular, consider a null hypothesis $H_0 : p = 0.80$ versus the alternative that the proportion of cups correctly filled exceeds 0.8, i.e. $H_a : p > 0.80$. We can use the normal approximation to the binomial for the computations. Consider an experiment where $n = 100$ cups are filled and assumed the null hypothesis was true (i.e. $p = 0.80$). A test with a significance level α close to 0.05 is desired. If we choose a cut-off value of $y = 87$ for the critical region and reject H_0 if $y \geq 87$, then the exact significance level α is

$$P(Y \geq 87) = 0.046912237$$

found using an exact binomial probability computation. A normal approximation can now be computed as:

$$\begin{aligned} P(Y \geq 87) &= 1 - P(Y < 87) \\ &= 1 - P(Y \leq 86) \\ &= 1 - P(Y \leq 86.5) \\ &= 1 - P\left(\frac{Y - np}{\sqrt{npq}} \leq \frac{86.5 - np}{\sqrt{npq}}\right) \\ &\approx 1 - P\left(Z \leq \frac{86.5 - np}{\sqrt{npq}}\right) \\ &= 1 - P\left(Z \leq \frac{86.5 - 100 \cdot 0.8}{\sqrt{100 \cdot 0.80 \cdot 0.20}}\right) \\ &= 1 - P(Z \leq 1.625) \\ &= 1 - 0.94791872 \\ &= 0.052081279, \end{aligned}$$

which is fairly close to the exact value.

Problems

- Let Y denote a binomial random variable with n trials and success probability $p = .2$. Find the exact value of $P(Y = 6)$ using the binomial probability formula and the approximate value using a normal approximation in each case below. In which cases does the normal approximation provide a reasonable answer?

a) $n = 10$

b) $n = 30$

2. In cup-a-soup example, consider a two-tailed test with rejection region: Reject H_0 if $Y \geq 88$ or if $Y \leq 72$ where Y was binomial on $n = 100$ trials. The null hypothesis was $H_0 : p = 0.80$. Compute the approximate significance level of this test using a normal approximation. That is, assuming $p = 0.80$, compute an approximation to $P(Y \geq 88) + P(Y \leq 72)$.

References

Albin, S. L. (1990), "The lognormal distribution for modeling quality data when the mean is near zero," *Journal of Quality Technology*, **22**, 105–110.

Flury, B. (1997), *A First Course in Multivariate Statistics*, New York: Springer.

Hurwitz, Arnon and Spagon, Pat (19xx). Statistical Methods for Test Procedures Sematech Report. Section 1.3.1.1

Lucas, J. M. (1985), "Counted data CUMSUM's," *Technometrics*, **27**, 129-144.