Chapter 6. Sampling and Estimation

6.1. Introduction

Frequently the engineer is unable to completely characterize the entire population. She/he must be satisfied with examining some subset of the population, or several subsets of the population, in order to infer information about the entire population. Such subsets are called **samples**. A **population** is the entirety of observations and a sample is a subset of the population. A sample that gives correct inferences about the population is a **random sample**, otherwise it is **biased**.

Statistics are given different symbols than the expectation values because *statistics are approximations of the expectation value*. The statistic called the mean is an approximation to the expectation value of the mean. The statistic mean is the mean of the sample and the expectation value mean is the mean of the entire population. In order to calculate an expectation, one requires knowledge of the PDF. In practice, the motivation in calculating a statistic is that one has no knowledge of the underlying PDF.

6.2. Statistics

Any function of the random variables constituting a random sample is called a statistic.

Example 6.1.: **Mean**

The mean is a statistic of a random sample of size *n* and is defined as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{6.1}$$

Example 6.2.: **Median**

The median is a statistic of a random sample of size *n*, which represents the “middle” value of the sample and, for a sampling arranged in increasing order of magnitude, is defined as

$$\tilde{X} = X_{(n+1)/2} \quad \text{for odd n}$$
$$\tilde{X} = \frac{X_{n/2} + X_{(n+1)/2}}{2} \quad \text{for even n} \tag{6.2}$$

The median of the sample space {1,2,3} is 2.
The median of the sample space {3,1,2} is 2.
The median of the sample space {1,2,3,4} is 2.5.

Example 6.3.: **Mode**

The mode is a statistic of a random sample of size *n*, which represents the most frequently appearing value in the sample. The mode may not exist and, if it does, it may not be unique.

The mode of the sample space {2,1,2,3} is 2.
The mode of the sample space {2,1,2,3,4,4} is 2 and 4. (bimodal)
The mode of the sample space {1,2,3} does not exist since each entry occurs only once.

Example 6.4.: **Range**

The range is a statistic of a random sample of size *n*, which represents the "span" of the sample and, for a sampling arranged in increasing order of magnitude, is defined as

$$range(X) = X_n - X_1 \tag{6.3}$$

The range of {1,2,3,4,5} is 5-1=4.

Example 6.5.: **Variance**

The variance is a statistic of a random sample of size *n*, which represents the "spread" of the sample and is defined as

$$S^2 = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2}{n-1} = \frac{n\sum_{i=1}^{n}\left(X_i\right)^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)} \tag{6.4}$$

The reason for using (*n*-1) in the denominator rather than *n* is given later.

Example 6.6.: **Standard Deviation**

The standard deviation, *s*, is a statistic of a random sample of size *n*, which represents the "spread" of the sample and is defined as the positive square root of the variance.

$$S = \sqrt{S^2} \tag{6.5}$$

6.3. Sampling Distributions

We have now stated the definitions of the statistics we are interested in. Now, we need to know the distribution of the statistics to determine how good these sampling approximations are to the true expectation values of the population.

Statistic 1. Mean when the variance is known: Sampling Distribution

If $\overline{X}$ is the mean of a random sample of size n taken from a population with mean μ and variance σ^2, then the limiting form of the distribution of

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \tag{6.6}$$

as $n \to \infty$, is the **standard normal distribution** $n(z;0,1)$. This is known as the Central Limit Theorem. What this says is that, given a collection of random samples, each of size *n*, yielding a mean $\overline{X}$, the distribution of $\overline{X}$ approximates a normal distribution, and becomes exactly a normal distribution as the sample size goes to infinity. The distribution of *X* does not have to be normal. Generally, the normal approximation for $\overline{X}$ is good if $n > 30$.

We provide a derivation in Appendix V proving that the distribution of the sample mean is given by the normal distribution.

Example 6.7.: distribution of the mean, variance known

In a reactor intended to grow crystals in solution, a "seed" is used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n = 10$. The population has variance in crystal size of $\sigma^2 = 1.0\ \mu m^2$. (We must know this from previous research.) The samples yield mean crystal sizes of $\bar{x} = 15.0$ μm. What is the likelihood that the true population mean, μ, is actually less than 14.0 μm?

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{15 - 14}{1 / \sqrt{10}} = 3.162$$

$$P(\mu < 14) = P(z > 3.162)$$

We have the change in sign because as μ increases, *z* decreases.

The evaluation of the cumulative normal probability distribution can be performed several ways. First, when the pioneers were crossing the plains in their covered wagons and they wanted to evaluate probabilities from the normal distribution, they used Tables of the cumulative normal PDF, such as those provided in the back of the statistics textbook. These tables are also available online. For example wikipedia has a table of cumulative standard numeral PDFs at

http://en.wikipedia.org/wiki/Standard_normal_table

Using the table, we find

$$P(\mu < 14) = P(z > 3.162) = 1 - P(z < 3.162) = 1 - 0.9992 = 0.0008$$

Second, we can use a modern computational tool like MATLAB to evaluate the probability. The problem can be worked in terms of the standard normal PDF (μ = 0 and σ = 1), which for $P(\mu < 14) = P(z > 3.162) = 1 - P(z < 3.162)$ is

```
>> p = 1 - cdf('normal',3.162,0,1)

p = 7.834478217108032e-04
```

Alternatively, the problem can be worked in terms of the non-standard normal PDF ($\bar{x} = 15$ and $\sigma/\sqrt{n} = 1/\sqrt{10}$), which for $P(\mu < 14)$

```
>> p = cdf('normal',14,15,1/sqrt(10))

p = 7.827011290012762e-04
```

The difference in these results is due to the round-off in 3.162, used as an argument in the function call for the standard normal distribution.

Based on our sampling data, the probability that the true sample mean is less than 14.0 μm is 0.078%.

Statistic 2. difference of means when the variance is known: Sampling Distribution

It is useful to know the sampling difference of two means when you want to determine whether there is a significant difference between two populations. This situation applies when you takes two random samples of size n_1 and n_2 from two different populations, with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2, respectively. Then the sampling distribution of the difference of means, $\bar{X}_1 - \bar{X}_2$, is approximately normal, distributed with mean

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2$$

and variance

$$\sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} - \frac{\sigma_2^2}{n_2}$$

Hence,

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} \tag{6.7}$$

is approximately a standard normal variable.

Example 6.8.: distribution of the difference of means, variances known

In a reactor intended to grow crystals, two different types of "seeds" are used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n_1 = 10$ and $n_2 = 20$. The populations have variances in crystal size of $\sigma_1^2 = 1.0$ μm^2 and $\sigma_2^2 = 2.0$ μm^2. (We must know this from previous research.) The samples yield mean crystal sizes of $\overline{X}_1 = 15.0$ μm and $\overline{X}_2 = 10.0$ μm. How confident can we be that the true difference in population means, $\mu_1 - \mu_2$, is actually 4.0 μm or greater?

Using equation (6.7) we have:

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} = \frac{(15-10)-(4)}{\sqrt{\left(\frac{1}{10}\right) + \left(\frac{2}{20}\right)}} = 2.2361$$

$$P(\mu_1 - \mu_2 > 4.0) = P(z < 2.2361)$$

We have the change in sign because as $\Delta\mu$ increases, *z* decreases. The probability that $\mu_1 - \mu_2$ is greater 4.0 μm is then given by *P*(*Z*<2.2361). How do we know that we want *P*(*Z*<2.2361) and not *P*(*Z*>2.2361)? We just have to sit down and think what the problem physically means. Since we want the probability that $\mu_1 - \mu_2$ is greater 4.0 μm, we know we need to include the area due to higher values of $\mu_1 - \mu_2$. Higher values of $\mu_1 - \mu_2$ yield lower values of *Z*. Therefore, we need the less than sign.

The evaluation of the cumulative normal probability distribution can again be performed two ways. First, using a standard normal table, we have

$$P(Z < 2.24) = 0.9875$$

Second, using MATLAB we have

```
>> p = cdf('normal',2.2361,0,1)
```

```
p =  0.987327389270190
```

We expect 98.73% of the differences in crystal size of the two populations to be at least 4.0 μm.

Statistic 3. Mean when the variance is unknown: Sampling Distribution

Of course, usually we don't know the population variance. In that case, we have to use some other statistic to get a handle on the distribution of the mean.

If $\overline{X}$ is the mean of a random sample of size *n* taken from a population with mean μ and unknown variance , then the limiting form of the distribution of

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \tag{6.8}$$

as $n \to \infty$, is the **t distribution** $f_T(t;v)$. The T-statistic has a t-distribution with *v*=*n*-1 degrees of freedom. The t-distribution is just another continuous PDF, like the others we learned about in the previous section.

The t distribution is given by

$$f(t) = \frac{\Gamma[(v+1)/2)]}{\Gamma(v/2)\sqrt{\pi v}}\left(1+\frac{t^2}{v}\right)^{-\frac{v+1}{2}} \quad \text{for} \quad -\infty < t < \infty$$

As a reminder, the t distribution is plotted again in Figure 6.1.

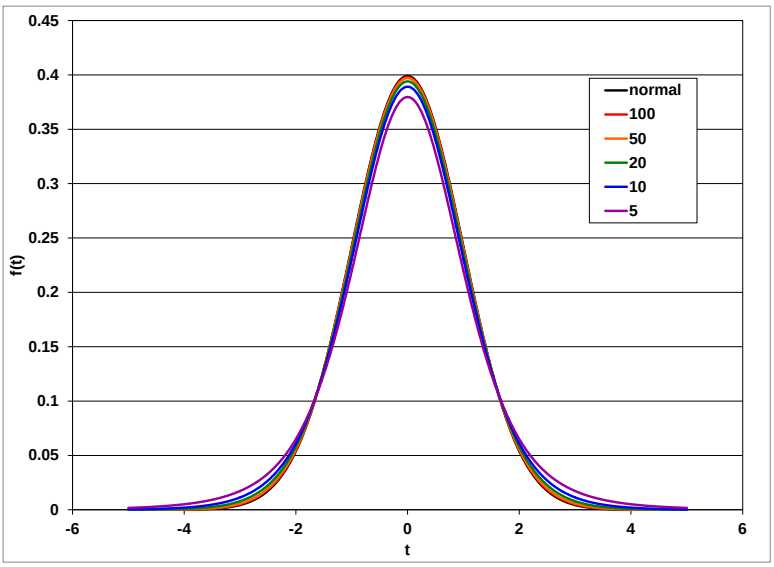

Figure 6.1. The t distribution as a function of the degrees of freedom and the normal distribution.

Example 6.9.: distribution of the mean, variance unknown

In a reactor intended to grow crystals, a "seed" is used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n = 10$. The population has unknown variance in crystal size. The samples yield mean crystal sizes of $\bar{x} = 15.0$ μm and a sample variance of $s^2 = 1.0\,\mu m^2$.

What is the likelihood that the true population mean , μ , is actually less than 14.0 μm?

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{15-14}{1/\sqrt{10}} = 3.162$$

$$P(\mu < 14) = P(t > 3.162)$$

We have the change in sign because as μ increases, *t* decreases. The parameter $v = n-1 = 9$.

The evaluation of the cumulative t probability distribution can again be performed two ways. First, we can use a table of critical values of the t-distribution. It is crucial to note that such a table does not provide cumulative PDFs, rather it provides one minus the cumulative PDF. In other words, where as the standard normal table provides the probability less than z (the cumulative PDF), the t-distribution table provides the probability greater than t (one minus the cumulative PDF). We then have

$$P(\mu < 14) = P(t > 3.162) \approx 0.007$$

Second, using MATLAB we have $P(\mu < 14) = P(t > 3.162) = 1 - P(t < 3.162)$

```
>> p = 1 - cdf('t',3.162,9)

p =   0.005756562560207
```

Based on our sampling data, the probability that the true sample mean is less than 14.0 μm is 0.57%.

We should point out that our percentage here is substantially greater than for our percentage when we knew the population variance (0.078%). That is because knowing the population variance reduces our uncertainty. Approximating the population variance with the sampling variance adds to the uncertainty and results in a larger percentage of our population deviating farther from the sample mean.

Example 6.10.: distribution of the mean, variance unknown

An engineer claims that the population mean yield of a batch process is 500 g/ml of raw material. To verify this, she samples 25 batches each month. One month the sample has a mean $\bar{X} = 518$ g and a standard deviation of s=40 g. Does this sample support his claim that $\mu = 500$ g?

The first step in solving this problem is to compute the T statistic.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{500-518}{40/\sqrt{25}} = -2.25$$

Second, using MATLAB we have $P(\mu > 518) = P(t < -2.25)$

```
>> p = cdf('t',-2.25,24)

p =  0.016944255452754
```

(Or using a Table, we find that when v=24 and T=2.25, α=0.02). This means there is only a 1.6% probability that a population with $\mu = 500$ would yield a sample with $\overline{X} = 518$ or higher. Therefore, it is unlikely that 500 is the population mean.

Statistic 4. difference of means when the variance is unknown: Sampling Distribution

It is useful to know the sampling difference of two means when you want to determine whether there is a significant difference between two populations. Sometimes you want to do this when you don't know the population variances. This situation applies when you takes two random samples of size n_1 and n_2 from two different populations, with means μ_1 and μ_2 and unknown variances. Then the sampling distribution of the difference of means, $\overline{X}_1 - \overline{X}_2$, follows the t-distribution.

transformation: $$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} \qquad (6.9)$$

symmetry: $t_{1-\alpha} = -t_\alpha$,

parameters: $v = n_1 + n_2 - 2$ if $\sigma_1 = \sigma_2$

parameters: $$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]}$$ if $\sigma_1 \neq \sigma_2$

Since we don't know either population variance in this case, we can't assume they are equal unless we are told they are equal.

Example 6.11.: distribution of the difference of means, variances unknown

In a reactor intended to grow crystals, two different types of "seeds" are used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n_1 = 10$ and $n_2 = 20$. The populations have unknown variances in crystal size. The samples yield

mean crystal sizes of $\overline{X}_1 = 15.0$ μm and $\overline{X}_2 = 10.0$ μm and sample variances of $s_1^2 = 1.0\,\mu m^2$ and $s_2^2 = 2.0\ \mu m^2$. What percentage of true population differences yielding these sampling results would have a true difference in population means, $\mu_1 - \mu_2$, of 4.0 μm or greater?

$$T = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}} = \frac{(15-10)-(4)}{\sqrt{\left(\frac{1}{10}\right) + \left(\frac{2}{20}\right)}} = 2.2361$$

The degree of freedom parameter is given by:

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]} = \frac{\left(\frac{1^2}{10} + \frac{2^2}{20}\right)^2}{\left[\left(\frac{1^2}{10}\right)^2 \Big/ (10 - 1)\right] + \left[\left(\frac{2^2}{20}\right)^2 \Big/ (20 - 1)\right]} = 27.98 \approx 28$$

$$P(\mu_1 - \mu_2 > 4.0) = P(t < 2.2361) = 1 - P(t > 2.2361)$$

The evaluation of the cumulative normal probability distribution can again be performed two ways. First, using a table of critical values of the t-distribution, we have

$$P(\mu_1 - \mu_2 > 4.0) = P(t < 2.2361) = 1 - P(t > 2.2361) = 1 - 0.0217 = 0.9783$$

Second, using MATLAB we have for $P(\mu_1 - \mu_2 > 4.0) = P(t < 2.2361)$

```
>>  p = cdf('t',2.2361,28)

p =  0.983252747598848
```

We expect 98.3% of the differences in crystal size of the two populations to be at least 4.0 μm.

Statistic 5. Variance: Sampling Distribution

We now wish to know the sampling distribution of the sample variance, S^2. If S^2 is the variance of a random sample of size n taken from a population with mean μ and variance σ^2, then the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \qquad (6.10)$$

has a chi-squared distribution with v=n-1 degrees of freedom, $f_{\chi^2}(\chi^2; n-1)$. The chi-squared distribution is defined as

$$f_{\chi^2}(x;v) = \begin{cases} \dfrac{1}{2^{v/2}\Gamma(v/2)} x^{v/2-1} e^{-x/2} & \text{for } x > 0 \\ 0 & \text{elsewhere} \end{cases}$$

It is a special case of the Gamma Distribution, when α=v/2 and β=2, where v is called the "degrees of freedom" and is a positive integer. As a reminder, we provide a plot of the chi-squared distribution in Figure 6.2.

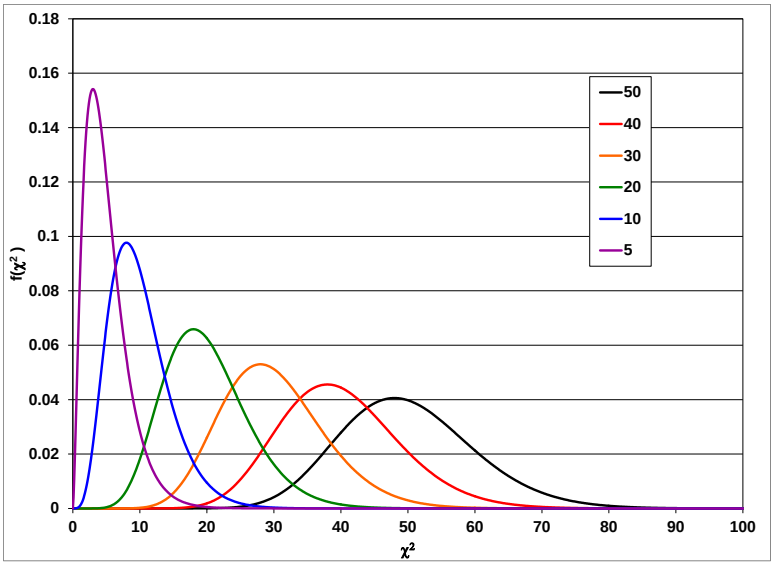

Figure 6.2. The chi-squared distribution for various values of v.

Example 6.12.: distribution of the variance

In a reactor intended to grow crystals, a "seed" is used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n = 10$. The samples yield mean crystal sizes of $\bar{x} = 15.0$ μm and a sample variance of $s^2 = 1.0\,\mu m^2$. What is the likelihood that the true population variance , σ^2, is actually less than 0.5 μm^2?

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \frac{(10-1)1}{0.5} = 18$$

$$P\left(\sigma^2 < 0.5\right) = P\left(\chi^2 > 18\right)$$

We have the change in sign because as σ^2 increases, χ^2 decreases. The parameter v = n-1 = 9.

The evaluation of the cumulative χ^2 probability distribution can again be performed two ways. First, we can use a table of critical values of the χ^2-distribution. It is crucial to note that such a table does not provide cumulative PDFs, rather it provides one minus the cumulative PDF. We then have

$$P\left(\sigma^2 < 0.5\right) = P\left(\chi^2 > 18\right) \approx 0.04$$

Second, using MATLAB we have $P\left(\sigma^2 < 0.5\right) = P\left(\chi^2 > 18\right) = 1 - P\left(\chi^2 < 18\right)$

```
>> p = 1 - cdf('chi2',18,9)

p =   0.035173539466985
```

Based on our sampling data, the probability that the true variance is less than 0.5 μm^2 is 3.5%.

Statistic 6. the ratio of 2 Variances: Sampling Distribution (F-distribution)

Just as we studied the distribution of two sample means, so too are we interested in the distribution of two variances. In the case of the mean, it was a difference. In the case of the variance, the ratio is more useful. Now consider sampling two random samples of size n_1 and n_2 from two different populations, with means σ_1^2 and σ_2^2, respectively. The statistic, F,

$$F = \frac{S_1^2 / \sigma_1^2}{S_2^2 / \sigma_2^2} = \frac{S_1^2 \sigma_2^2}{S_2^2 \sigma_1^2} \tag{6.11}$$

provides a distribution of the ratio of two variances. This distribution is called the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. The f-distribution is defined as

$$h_f(f; v_1, v_2) = \begin{cases} \dfrac{\Gamma\left(\dfrac{v_1 + v_2}{2}\right)\left(\dfrac{v_1}{v_2}\right)^{\frac{v_1}{2}}}{\Gamma\left(\dfrac{v_1}{2}\right)\Gamma\left(\dfrac{v_2}{2}\right)} \dfrac{f^{\frac{v_1}{2}-1}}{\left(1 + \dfrac{v_1}{v_2} f\right)^{\frac{v_1+v_2}{2}}} & \text{for } f > 0 \\ 0 & \text{elsewhere} \end{cases}$$

As a reminder, the f-distribution is plotted in Figure 6.3.

Example 6.13.: ratio of the variances

In a reactor intended to grow crystals, two different types of "seeds" are used to encourage nucleation. Individual crystals are randomly sampled from the effluent of each reactor of sizes $n_1 = 10$ and $n_2 = 20$. The populations have unknown variances in crystal size. The samples yield

mean crystal sizes of $\overline{X}_1 = 15.0\ \mu m$ and $\overline{X}_2 = 10.0\ \mu m$ and sample variances of $s_1^2 = 1.0\,\mu m^2$ and $s_2^2 = 2.0\ \mu m^2$. What is the probability that the ratio of variances, $\frac{\sigma_1^2}{\sigma_2^2}$, is less than 0.25?

$$F = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} = \frac{1}{2 \cdot 0.25} = 2$$

$$P\left(\frac{\sigma_1^2}{\sigma_2^2} < 0.25\right) = P(F > 2)$$

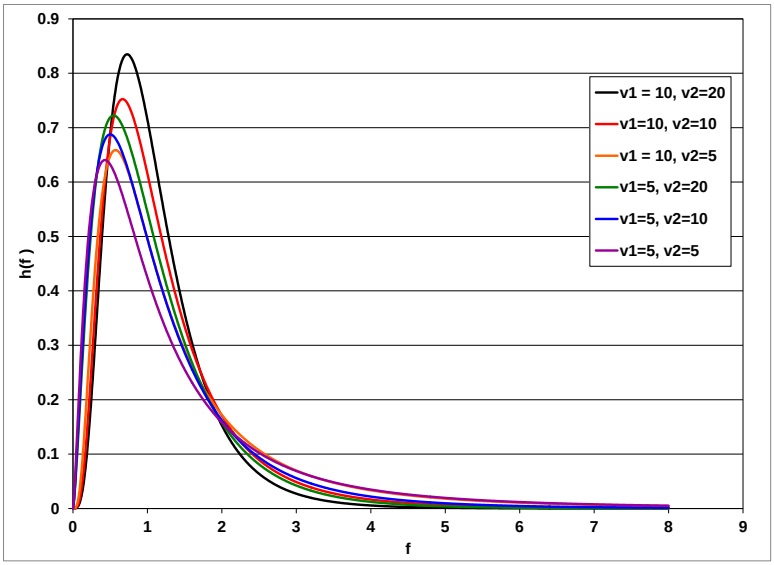

Figure 6.3. The F distribution for various values of v_1 and v_2.

We have the change in sign because as $\frac{\sigma_1^2}{\sigma_2^2}$ increases, F decreases. The parameters are $v_1 = n_1 - 1 = 9$ and $v_2 = n_2 - 1 = 19$.

The evaluation of the cumulative F probability distribution can again be performed in one way. We cannot use tables because there are no tables for arbitrary values of the probability. There are only tables for two values of the probability, 0.01 and 0.05. Therefore, using MATLAB we have $P\left(\frac{\sigma_1^2}{\sigma_2^2} < 0.25\right) = P(F > 2) = 1 - P(F < 2)$

```
>>  p = 1 - cdf('f',2,9,19)

p =  0.097413204997132
```

Based on our sampling data, the probability that the ratio of variances is less than 0.25 is 9.7%.

6.4. Confidence Intervals

In the previous section we showed what types of distributions describe various statistics of a random sample. In this section, we discuss estimating the population mean and variance from the sample mean and variance. In addition, we introduce confidence intervals to quantify the goodness of these estimates.

A confidence interval is some subset of random variable space with which someone can say something like, "I am 95% sure that the true population mean is between μ_{low} and μ_{hi}." In this section, we discuss how a confidence interval is defined and calculated.

The confidence interval is defined by a percent. This percent is called (1-2α). So if α=0.05, then you would have a 90% confidence interval.

The concept of a confidence interval is illustrated in graphical terms in Figure 6.4.

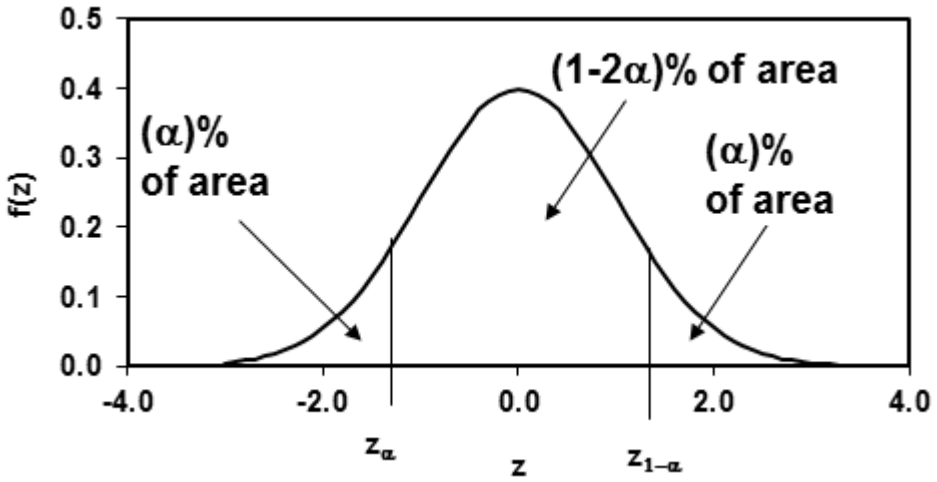

Figure 6.4. A schematic illustrating a confidence interval.

The trick then is to find $\mu_{low} = z_\alpha$ and $\mu_{hi} == z_{1-\alpha}$ so that you can say for a given α, I am $(1-2\alpha)\%$ confident that $\mu_{low} < \mu < \mu_{hi}$.

Statistic 1. mean, σ known: confidence interval

We now know that the sample mean is distributed with the standard normal distribution. For a symmetric PDF, centered around zero, like the standard normal, $\mu_{low} = -\mu_{hi}$. We can then make the statement:

$$P(z_\alpha < Z < z_{1-\alpha}) = 1 - 2\alpha$$

Now the normal distribution is symmetric about the y-axis so we can write

$$z_\alpha = -z_{1-\alpha}$$

so

$$P(z_\alpha < Z < z_{1-\alpha}) = P(z_\alpha < Z < -z_\alpha) = 1 - 2\alpha$$

where

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}.$$

We can rearrange this to equation to read

$$P(\overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}) = 1 - 2\alpha \qquad (6.12)$$

where we now have μ_{low} and μ_{hi} explicitly.

Example 6.14.: confidence interval on mean, variance known

Samples of dioxin contamination in 36 front yards in St. Louis show a concentration of 6 ppm. Find the 95% confidence interval for the population mean. Assume that the standard deviation is 1.0 ppm.

To solve this, first calculate $\alpha, z_\alpha, z_{1-\alpha}$.

$$1 - 2\alpha = 0.95$$
$$\alpha = 0.025$$
$$z_\alpha = z_{0.025} = -1.96$$
$$z_{1-\alpha} = -z_\alpha = 1.96$$

The z value came from a standard normal table. Alternatively, we can compute this value from MATLAB,

```
>> z = icdf('normal',0.025,0,1)

z =  -1.959963984540055
```

Here we used the inverse cumulative distribution function (icdf) command. Since we have the standard normal PDF, the mean is 0 and the variance is 1. The value of 0.025 corresponds to alpha, the probability.

To get the value of the other limit, we either rely on symmetry, or compute it directly,

```
>> z = icdf('normal',0.975,0,1)

z =  1.959963984540054
```

Note that these values of z are independent of all aspects of the problem except the value of the confidence interval.

Therefore, by equation (6.12)

$$P(6+(-1.96)\frac{1}{\sqrt{36}}<\mu<\overline{X}-(-1.96)\frac{1}{\sqrt{36}}=1-0.05=0.95$$

so the 95% confidence interval for the mean is $5.673<\mu<6.327$.

Statistic 2. mean, σ unknown: confidence interval

Now usually, we don't know the variance. We have to use our estimate of the variance, *s*, for σ. In that case, estimating the mean requires the T-distribution. (See previous section.) Let me stress that we do everything exactly as we did before but we use *s* for σ and use the *t*-distribution instead of the normal distribution. Remember the t-distribution is also symmetric about the origin, so $t_{1-\alpha}=-t_{\alpha}$. (this means you only have to compute the *t* probability once. Remember, *v*=*n*-1.

$$P(t_{\alpha}<T<t_{1-\alpha})=P(t_{\alpha}<T<-t_{\alpha})=1-2\alpha$$

where

$$T=\frac{\overline{X}-\mu}{s/\sqrt{n}}.$$

Just as before, we can rearrange this to equation to read

$$P(\overline{X}+t_{\alpha}\frac{s}{\sqrt{n}}<\mu<\overline{X}-t_{\alpha}\frac{s}{\sqrt{n}})=1-2\alpha \qquad (6.13)$$

where we now have μ_{low} and μ_{hi} explicitly.

Example 6.15.: confidence interval on mean, variance unknown

Samples of dioxin contamination in 36 front yards in St. Louis show a concentration of 6 ppm. Find the 95% confidence interval for the population mean. The sample standard deviation, *s*, was measured to be 1.0.

To solve this, first calculate $\alpha, t_{\alpha}, t_{1-\alpha}$ for v = 35.

$1-2\alpha=0.95$

$\alpha=0.025$

$t_{\alpha}=t_{0.025}=-2.03$

$t_{1-\alpha}=-t_{\alpha}=+2.03$

The t value came from a table of t-distribution values. Alternatively, we can compute this value using MATLAB,

```
>> t = icdf('t',0.025,35)

t =  -2.030107928250342
```

and for the upper limit

```
>> t = icdf('t',0.975,35)

t =  2.030107928250342,
```

which can also be obtained by symmetry. Note that these values of t are independent of all aspects of the problem except the value of the confidence interval and the number of sample points, n.

Therefore, by equation (6.13)

$$P(6-(2.03)\frac{1}{\sqrt{36}}<\mu<\overline{X}+(2.03)\frac{1}{\sqrt{36}}=1-0.05=0.95$$

so the 95% confidence interval for the mean is $5.662<\mu<6.338$.

You should note that we are a little less confident about the mean when we use the sample variance as the estimate for the population variance, for which the 95% confidence interval for the mean was $5.673<\mu<6.327$.

Statistic 3. difference of means, σ known: confidence interval

The exact same derivation that we used above for a single mean can be used for the difference of means. When we the variances of the two samples are known, we have:

$$P\left[\left(\overline{X}_1-\overline{X}_2\right)+z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\left(\mu_1-\mu_2\right)<\left(\overline{X}_1-\overline{X}_2\right)-z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]=1-2\alpha \qquad (6.14)$$

where z is a random variable obeying the standard normal PDF.

Example 6.16.: confidence interval on the difference of means, variances known

Samples of dioxin contamination in 36 front yards in Times Beach, a suburb of St. Louis, show a concentration of 6 ppm with a population variance of 1.0 ppm^2. Samples of dioxin contamination in 16 front yards in Quail Run, another suburb of St. Louis, show a concentration of 8 ppm with a population variance of 3.0 ppm^2. Find the 95% confidence interval for the difference of population means. .

To solve this, first calculate $\alpha, z_\alpha, z_{1-\alpha}$.

$1-2\alpha = 0.95$

$\alpha = 0.025$

$z_\alpha = z_{0.025} = -1.96$

$z_{1-\alpha} = -z_\alpha = 1.96$

The z value came from a table of standard normal PDF values. Alternatively, we can compute this value from MATLAB,

```
>> z = icdf('normal',0.025,0,1)

z =  -1.959963984540055
```

Therefore, by equation (6.16)

$$P\left[(6-8)-1.96\sqrt{\frac{1}{36}+\frac{3}{16}} < (\mu_1-\mu_2) < (6-8)+1.96\sqrt{\frac{1}{36}+\frac{3}{16}}\right] = 1-2(0.025)$$

$$P[-2.909 < (\mu_1-\mu_2) < -1.091] = 0.95$$

So the 95% confidence interval for the mean is $-2.909 < (\mu_1-\mu_2) < -1.091$.

If we are determining which site is more contaminated, then we are 95% sure that site 2 (Quail Run) is more contaminated by 1 to 3 ppm than site 1, (Times Beach).

Statistic 4. difference of means, σ unknown: confidence interval

When we the variances of the two samples are unknown, we have:

$$P\left[(\bar{X}_1-\bar{X}_2)+t_\alpha\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}} < (\mu_1-\mu_2) < (\bar{X}_1-\bar{X}_2)+t_{1-\alpha}\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right] = 1-2\alpha \qquad (6.15)$$

where the number of degrees of freedom for the t-distribution is

$v = n_1 + n_2 - 2$ if $\sigma_1 = \sigma_2$

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]} \text{ if } \sigma_1 \neq \sigma_2$$

Example 6.16.: confidence interval on the difference of means, variances unknown

Samples of dioxin contamination in 36 front yards in Times Beach, a suburb of St. Louis, show a concentration of 6 ppm with a sample variance of 1.0 ppm^2. Samples of dioxin contamination in 16 front yards in Quail Run, another suburb of St. Louis, show a concentration of 8 ppm with a sample variance of 3.0 ppm^2. Find the 95% confidence interval for the difference of population means. .

To solve this, first calculate $\alpha, t_\alpha, t_{1-\alpha}$.

$$v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]} = \frac{\left(\frac{1}{36} + \frac{3}{16}\right)^2}{\left[\left(\frac{1}{36}\right)^2 \Big/ (36 - 1)\right] + \left[\left(\frac{3}{16}\right)^2 \Big/ (16 - 1)\right]} = 19.59 \approx 20$$

$1 - 2\alpha = 0.95$

$\alpha = 0.025$

$t_\alpha = t_{0.025} = 2.086$

$t_{1-\alpha} = -t_\alpha = -2.086$

The t value came from a table of t-PDF values. Alternatively, we can compute this value using MATLAB,

```
>> t = icdf('t',0.025,20)

t =  -2.085963447265864
```

Therefore, substituting into equation (6.15) yields

$$P\left[(6-8) - 2.086\sqrt{\frac{1}{36} + \frac{3}{16}} < (\mu_1 - \mu_2) < (6-8) + 2.086\sqrt{\frac{1}{36} + \frac{3}{16}}\right] = 1 - 2(0.025)$$

$$P[-2.97 < (\mu_1 - \mu_2) < -1.03] = 0.95$$

So the 95% confidence interval for the mean is $-2.97 < (\mu_1 - \mu_2) < -1.03$.

If we are determining which site is more contaminated, then we are 95% sure that site 2 (Quail Run) is more contaminated by 1 to 3 ppm than site 1, (Times Beach).

Statistic 5. variance: confidence interval

The confidence interval of the variance can be estimated in a precisely analogous way, knowing that the statistic

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2}$$

has a chi-squared distribution with v=n-1 degrees of freedom, $f_{\chi^2}(\chi^2; n-1)$. So

$$P\left[\frac{(n-1)s^2}{\chi^2_{1-\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{\alpha}}\right] = 1-2\alpha \qquad (6.16)$$

Perversely, the tables of the critical values for the χ^2 distribution, have defined α to be 1-α, so the indices have to be switched when using the table.

$$P\left[\frac{(n-1)s^2}{\chi^2_{\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha}}\right] = 1-2\alpha$$ when using the χ^2 critical values table only!

If you get confused, just remember that the upper limit must be greater than the lower limit. Remember also that the $f_{\chi^2}(\chi^2; n-1)$ is not symmetric about the origin, so we cannot use the symmetry arguments used for the confidence intervals for functions of the mean.

Example 6.17.: variance

Samples of dioxin contamination in 16 front yards in St. Louis show a concentration of 6 ppm. Find the 95% confidence interval for the population mean. The sample standard deviation, *s*, was measured to be 1.0.

To solve this, first calculate $\alpha, \chi^2_{\alpha}, \chi^2_{1-\alpha}$.

For v = *n* – 1 = 15, we have

$1-2\alpha = 0.95$

$\alpha = 0.025$

$\chi^2_{\alpha} = \chi^2_{0.025} = 27.488$

$\chi^2_{1-\alpha} = \chi^2_{0.975} = 6.262$

The t value came from a table of χ^2-distribution values. Alternatively, we can compute this value using MATLAB,

```
>>  chi2 = icdf('chi2',0.025,15)

chi2 =   6.262137795043251
```

and

```
>>  chi2 = icdf('chi2',0.975,15)

chi2 =  27.488392863442972
```

Therefore, substituting into equation (6.16) yields

$$P\left[\frac{(16-1)1.0}{27.488} < \sigma^2 < \frac{(16-1)1.0}{6.262}\right] = 1 - 2(0.025)$$

$$P\left(0.5457 < \sigma^2 < 2.395\right) = 0.95$$

So the 95% confidence interval for the mean is $0.5457 < \sigma^2 < 2.395$.

Statistic 6. ratio of variances: confidence interval (p. 253)

The ratio of two population variances can be estimated in a precisely analogous way, knowing that the statistic

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2}$$

follows the F-distribution with $v_1 = n_1 - 1$ and $v_2 = n_2 - 1$ degrees of freedom. Remember, the F-distribution has a symmetry, $f_{1-\alpha/2}(v_1, v_2) = \dfrac{1}{f_{\alpha/2}(v_2, v_1)}$. This symmetry relation is essential if one is to use tables for the critical value of the F-distribution. It is not essential if one uses MATLAB commands.

If one is computing the cumulative PDF for the f distribution, then one simply, rearranges this equation for $\dfrac{\sigma_1^2}{\sigma_2^2}$

$$\frac{\sigma_2^2}{\sigma_1^2} = F\frac{S_2^2}{S_1^2}$$

$$\frac{\sigma_1^2}{\sigma_2^2} = \frac{1}{F}\frac{S_1^2}{S_2^2}$$

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{1-\alpha}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha}(v_1,v_2)}\right] = 1-2\alpha \qquad (6.17)$$

One notes that the order of the limits has changed here, since as $\frac{\sigma_1^2}{\sigma_2^2}$ goes up, F goes down. In any case, the lower limit must be smaller than the upper limit. If one chooses to use tables of critical values, one must take into account two idiosyncrasies of the procedure. First, as was the case with the t and chi-squared distributions, the table provide the probability that f is greater than a value, not the cumulative PDF, which is the probability that f is less than a value. Second, the tables only provide data for small values of α. Therefore, we must eliminate all instances of 1-α., using a symmetry relation. The result is

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_{\alpha}(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_{\alpha}(v_2,v_1)\right] = 1-2\alpha \quad \text{when using the tables only!}$$

Example 6.18.: confidence interval on the ratio of variances

Samples of dioxin contamination in 20 front yards in Times Beach, a suburb of St. Louis, show a concentration of 6 ppm with a sample variance of 1.0 ppm^2. Samples of dioxin contamination in 16 front yards in Quail Run, another suburb of St. Louis, show a concentration of 8 ppm with a sample variance of 3.0 ppm^2. Find the 90% confidence interval for the difference of population means. .

To solve this, first calculate $\alpha, F_{\alpha}, F_{1-\alpha}$, with $v_1 = n_1 - 1 = 19$ and $v_2 = n_2 - 1 = 15$

$1-2\alpha = 0.90$

$\alpha = 0.05$

We can compute the f probabilities using MATLAB,

```
>>  f = icdf('f',0.05,19,15)

f =   0.447614966503185
```

and

```
>> f = icdf('f',0.95,19,15)

f =   2.339819281665456
```

Substituting into equation (6.16) yields

$$P\left[\frac{1}{3}\frac{1}{2.3398} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{3}\frac{1}{0.4476}\right] = 1 - 2(0.05)$$

$$P\left[0.1425 < \frac{\sigma_1^2}{\sigma_2^2} < 0.7447\right] = 0.90$$

Alternatively, we can use the table of critical values

$$F_\alpha = F_{0.05} = F_{0.05}(v_1 = 19, v_2 = 15) \approx F_{0.05}(v_1 = 20, v_2 = 15) = 2.33$$
$$F_{0.05}(v_1 = 15, v_2 = 19) = 2.23$$

$$P\left[\frac{1}{3}\frac{1}{2.33} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{1}{3}2.23\right] = 1 - 2(0.05)$$

$$P\left[0.1431 < \frac{\sigma_1^2}{\sigma_2^2} < 0.7433\right] = 0.90$$

So the 90% confidence interval for the mean is $0.1425 < \frac{\sigma_1^2}{\sigma_2^2} < 0.7447$.

If we are determining which site has a greater variance of contamination levels then we are 90% sure that site 2 (Quail Run) has more variance by a factor of 1.3 to 7.0.

6.5. Problems

We intend to purchase a liquid as a raw material for a material we are designing. Two vendors offer us samples of their product and a statistic sheet. We run the samples in our own labs and come up with the following data:

Vendor 1		Vendor 2	
sample #	outcome	sample #	outcome
1	2.3	1	2.49
2	2.49	2	1.98
3	2.05	3	2.18
4	2.4	4	2.36
5	2.18	5	2.47
6	2.12	6	2.36
7	2.38	7	1.82
8	2.39	8	1.88
9	2.4	9	1.87
10	2.46	10	1.87
11	2.19		
12	2.04		
13	2.43		
14	2.34		
15	2.19		
16	2.12		

Vendor Specification Claims:

Vendor 1: $\mu = 2.0$ and $\sigma^2 = 0.05$, $\sigma = 0.2236$

Vendor 2: $\mu = 2.3$ and $\sigma^2 = 0.12$, $\sigma = 0.3464$

Sample statistics, based on the data provided in the table above.

$$n_1 = 16 \qquad \bar{x}_1 = \frac{1}{16}\sum_{i=1}^{16} x_i = 2.280 \qquad s_1^{\,2} = \frac{1}{16}\sum_{i=1}^{16}\left[(x_i - \bar{x}_1)^2\right] = 0.0229 \qquad s_1 = 0.1513$$

$$n_2 = 10 \qquad \bar{x}_2 = \frac{1}{10}\sum_{i=1}^{10} x_i = 2.128 \qquad s_2^{\,2} = \frac{1}{10}\sum_{i=1}^{10}\left[(x_i - \bar{x}_2)^2\right] = 0.0744 \qquad s_2 = 0.2728$$

Problem 6.1.

Determine a 95% confidence interval on the mean of sample 1. Use the value of the population variance given. Is the given population mean legitimate?

Problem 6.2.

Determine a 95% confidence interval on the difference of means between samples 1 and 2. Use the values of the population variance given. Is the difference between the given population means legitimate?

Problem 6.3.

Determine a 95% confidence interval on the mean of sample 1. Assume the given values of the population variances are suspect and not to be trusted. Is the given population mean legitimate?

Problem 6.4.

Determine a 95% confidence interval on the difference of means between samples 1 and 2. Assume the given values of the population variances are suspect and not to be trusted. Is the difference between the given population means legitimate?

Problem 6.5.

Determine a 95% confidence interval on the variance of sample 1. Is the given population variance legitimate?

Problem 6.6.

Determine a 98% confidence interval on the ratio of variance of samples 1 & 2. Is the ratio of the given population variances legitimate?