Lectures 19 - Sampling and Estimation

Text: WMM, Chapters 8-9. Sections 8.1, 8.2, 8.5-8.8, 9.3-9.4,9.7,9.11-9.12

19.1 Samples (WMM, p. 198)

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. Got it?

19.2 Statistics (WMM, 201)

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

Example 19.1: **Mean**

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

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (19.1)$$

Example 19.2: **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

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

$$\tilde{X} = \frac{X_{n/2} + X_{(n+1)/2}}{2} \quad \text{for even } n \qquad (19.2)$$

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 19.3: **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 19.4: **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

$$\text{range}(X) = X_n - X_1 \tag{19.3}$$

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

Example 19.5: **Variance**

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

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} = \frac{n\sum_{i=1}^{n}(X_i)^2 - \left(\sum_{i=1}^{n} X_i\right)^2}{n(n-1)} \tag{19.4}$$

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

Example 19.6: **Standard Deviation**

The standard deviation, σ, 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.

19.3 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 (WMM, p. 217)

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

$$Z = \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \tag{19.5}$$

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.

Derivation of the fact that the distribution of the mean is the normal distribution

Consider taking n samples from a population characterized by mean, μ, and variance, σ^2. The sample mean is given by $\bar{x}$.

We define a moment generating function for a continuous PDF to be:

$$M_x(t) = \mu_{e^{tx}} = E\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

From this definition, and using the rules of linear operation, we can show that

$$M_{x+a}(t) = \mu_{e^{t(x+a)}} = E\left[e^{t(x+a)}\right] = \int_{-\infty}^{\infty} e^{t(x+a)} f(x)dx = \int_{-\infty}^{\infty} e^{tx} e^{ta} f(x)dx = e^{ta} \int_{-\infty}^{\infty} e^{tx} f(x)dx$$

$$M_{x+a}(t) = e^{ta} M_x(t) \tag{1}$$

$$M_{ax}(t) = \mu_{e^{tax}} = E\left[e^{tax}\right] = \int_{-\infty}^{\infty} e^{tax} f(x)dx = M_x(at)$$

$$M_{ax}(t) = M_x(at) \tag{2}$$

$$M_{x_1+x_2+x_3+\ldots+x_n}(t) = \mu_{e^{t(x_1+x_2+x_3+\ldots+x_n)}} = E\left[e^{t(x_1+x_2+x_3+\ldots+x_n)}\right] = \int_{-\infty}^{\infty} e^{t(x_1+x_2+x_3+\ldots+x_n)} f(x)dx$$

$$M_{x_1+x_2+x_3+\ldots+x_n}(t) = M_{x_1}(t) M_{x_2}(t) M_{x_3}(t) \ldots M_{x_n}(t) \tag{3}$$

So that

$$M_{(\bar{x}-\mu)/(\sigma/\sqrt{n})}(t) = \mu_{e^{t\left[(\bar{x}-\mu)/(\sigma/\sqrt{n})\right]}} = E\left[e^{t\left[(\bar{x}-\mu)/(\sigma/\sqrt{n})\right]}\right] = \int_{-\infty}^{\infty} e^{t\left[(\bar{x}-\mu)/(\sigma/\sqrt{n})\right]} f(\bar{x})d\bar{x}$$

$$M_{(\bar{x}-\mu)/(\sigma/\sqrt{n})}(t) = \int_{-\infty}^{\infty} e^{t\bar{x}/(\sigma/\sqrt{n})} e^{-t\mu/(\sigma/\sqrt{n})} f(\bar{x})d\bar{x} = e^{-t\mu\sqrt{n}/\sigma} \int_{-\infty}^{\infty} e^{t\bar{x}\sqrt{n}/\sigma} f(\bar{x})d\bar{x}$$

$$M_{(\bar{x}-\mu)/(\sigma/\sqrt{n})}(t) = e^{-t\mu\sqrt{n}/\sigma} M_{\bar{x}}\left(\frac{t\sqrt{n}}{\sigma}\right)$$

Now consider that $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, so

$$M_{\bar{x}}\left(\frac{t\sqrt{n}}{\sigma}\right) = M_{\frac{1}{n}\sum_{i=1}^{n} X_i}\left(\frac{t\sqrt{n}}{\sigma}\right) = \int_{-\infty}^{\infty} e^{t\sqrt{n}/\sigma \frac{1}{n}\sum_{i=1}^{n} X_i} f(x)dx = \int_{-\infty}^{\infty} e^{t/(\sigma\sqrt{n})\sum_{i=1}^{n} X_i} f(x)dx$$

$$M_{\bar{x}}(t\sqrt{n}/\sigma) = \int_{-\infty}^{\infty} e^{t/(\sigma\sqrt{n})\sum_{i=1}^{n} X_i} f(x)dx = M_{x_1}\left(\frac{t}{\sigma\sqrt{n}}\right) M_{x_2}\left(\frac{t}{\sigma\sqrt{n}}\right)..M_{x_n}\left(\frac{t}{\sigma\sqrt{n}}\right)$$

Since there isn't anything intrinsic that distinguishes one X_i from another, we can write

$$M_{\bar{x}}(t\sqrt{n}/\sigma) = \left[M_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

If we substitute this back into our original equation

$$M_{\frac{(\bar{x}-\mu)}{(\sigma/\sqrt{n})}}(t) = e^{-t\mu\sqrt{n}/\sigma} M_{\bar{x}}\left(\frac{t\sqrt{n}}{\sigma}\right) = e^{-t\mu\sqrt{n}/\sigma}\left[M_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right]^n$$

Take the natural log of both sides:

$$\ln\left[M_{\frac{(\bar{x}-\mu)}{(\sigma/\sqrt{n})}}(t)\right] = \frac{-t\mu\sqrt{n}}{\sigma} + n\ln\left[M_x\left(\frac{t}{\sigma\sqrt{n}}\right)\right]$$

Expand $M_x\left(\frac{t}{\sigma\sqrt{n}}\right)$ as an infinite series in powers of t about t=0.

$$M_x\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + v_1 t + v_2\frac{t^2}{2!} + v_3\frac{t^3}{3!} + \ldots + v_r\frac{t^r}{r!} + \ldots$$

where

$$v_i = \left.\frac{d^i M_x\left(\frac{t}{\sigma\sqrt{n}}\right)}{dt^i}\right|_{t=0}$$

We can write this as

$$M_x\left(\frac{t}{\sigma\sqrt{n}}\right) = 1 + v(t)$$

where v(t) is an infinite series in t. For very large sample sizes, n

$$\lim_{n\to\infty} \ln\left[M_{\frac{(\bar{x}-\mu)}{(\sigma/\sqrt{n})}}(t)\right] = \lim_{n\to\infty} \ln[1+v(t)] = \frac{t^2}{2}$$

This can be shown by expanding the natural log in a Mclaurin series. For the present purposes, we will take this step given above on faith. Then, we have

$$\lim_{n\to\infty} M_{\frac{(\bar{x}-\mu)}{(\sigma/\sqrt{n})}}(t) = e^{\frac{t^2}{2}}$$

So the first moment of $Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ in the limit of large n is $e^{\frac{t^2}{2}}$

Well, let's find what the moment of the random variable, z, would be if it follows the normal distribution. The PDF of the normal distribution is

$$f(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$M_x(t) = \mu_{e^{tx}} = E\left[e^{tx}\right] = \int_{-\infty}^{\infty} e^{tx} f(x)dx = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} dx$$

$$M_x(t) = \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2+2x\mu-\mu^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} e^{\frac{2tx\sigma^2}{2\sigma^2}} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2+2x\mu-\mu^2}{2\sigma^2}} dx$$

$$M_x(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{2tx\sigma^2-x^2+2x\mu-\mu^2}{2\sigma^2}} dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-x^2+2\left(t\sigma^2+\mu\right)x-\mu^2}{2\sigma^2}} dx$$

Complete the square in the exponent:

$$-x^2 + 2\left(t\sigma^2+\mu\right)x - \mu^2 = \left[x-\left(t\sigma^2+\mu\right)\right]^2 - 2\mu t\sigma^2 - t^2\sigma^4$$

$$M_x(t) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\left[x-\left(t\sigma^2+\mu\right)\right]^2 - 2\mu t\sigma^2 - t^2\sigma^4}{2\sigma^2}} dx = e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{\left[x-\left(t\sigma^2+\mu\right)\right]^2}{2\sigma^2}} dx$$

Let $w = \frac{\left[x-\left(t\sigma^2+\mu\right)\right]}{\sigma}$ so that $dw = \frac{dx}{\sigma}$ and

$$M_x(t) = e^{\mu t + \frac{t^2\sigma^2}{2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{w^2}{2}} dw = e^{\mu t + \frac{t^2\sigma^2}{2}} (1) = e^{\mu t + \frac{t^2\sigma^2}{2}}$$

So that the first moment generating function of the standard normal PDF is

$$M_x(t) = e^{\frac{t^2}{2}}$$

If we compare this moment generating function with that obtained for

$$\lim_{n\to\infty} M_{\frac{(\bar{x}-\mu)}{(\sigma/\sqrt{n})}}(t) = e^{\frac{t^2}{2}}$$

we find that they are the same in the limit of large n. Since there is a one-to-one correspondence between PDFs and moment-generating functions, we see that the PDF for $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$ is the standard normal PDF.

Example: distribution of the mean, variance known

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 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.

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

Based on our sampling data, the probability that the true sample mean is less than 14.0 μm is 0.08%.

Statistic 2. difference of means when the variance is known: Sampling Distribution (WMM, p. 220)

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, $\overline{X}_1 - \overline{X}_2$, is approximately normal, distributed with mean

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

and variance

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

Hence,

$$Z = \frac{\left(\overline{X}_1 - \overline{X}_2\right) - \left(\mu_1 - \mu_2\right)}{\sqrt{\left(\frac{\sigma_1^2}{n_1}\right) + \left(\frac{\sigma_2^2}{n_2}\right)}} \qquad (19.6)$$

is approximately a standard normal variable.

Example: 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 $\sigma_1^2 = 1.0\,\mu m^2$ and $\sigma_2^2 = 2.0\ \mu m^2$. (We must know this from previous research.) The samples yield mean crystal sizes of $\overline{X}_1 = 15.0\ \mu m$ and $\overline{X}_2 = 10.0\ \mu m$. 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 (19.6) we have:

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

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

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, which from Table A.3 yields:

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

We expect 98.75% 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 $\bar{X}$ 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

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

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 DPF, like the others we learned about in the previous section. The t-distribution is defined on p. 229 of WMM but we do not derive or reprint it here since we are engineers who just want to know about the application of statistics. Values of the t-distribution are tabulated in Table A.4.

NOTE: In Table A.4, you are given $P(t > t_\alpha) = \alpha$ and not $P(t < t_\alpha) = \alpha$, as you were with the normal distribution table.

Example: 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.

$$P(\mu < 14) = P(t > 3.162) \approx 0.001$$ from Table A.4

Based on our sampling data, the probability that the true sample mean is less than 14.0 μm is 0.1%.

We should point out that our percentage here is slightly greater than for our percentage when we knew the population variance. 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 slightly larger percentage of our population deviating farther from the sample mean.

Example #2: distribution of the mean, variance unknown

A chemical engineers claims that the population mean yield of a batch process is 500 g pr 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?

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

Using Table A.4, we find that when v=24 and T=2.25, α=0.02. This means there is only a 2% probability that a population with $\mu = 500$ would yield a sample with $\bar{X} = 518$. Therefore, it is unlikely that 500 is the population mean. In fact the yield is probably better than the engineer claimed. (Notice when we use Table A.5, we have the tabulated value, T, and find one of the table indices, α. Some problems would give, v and α, prompting you to find the T value.)

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

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

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

parameters: $$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left[\left(\frac{s_1^2}{n_1}\right)^2 \Big/ (n_1 - 1)\right] + \left[\left(\frac{s_2^2}{n_2}\right)^2 \Big/ (n_2 - 1)\right]}$$ 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: 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. (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 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:

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

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

We have the change in sign because as $\Delta\mu$ increases, z decreases. We expect 97.83% of the differences in crystal size of the two populations to be at least 4.0 μm.

Statistic 5. 19.3.3. Variance: Sampling Distribution WMM (p. 224)

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

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{\sigma^2} \qquad (19.7)$$

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

The probability that a random sample produces a χ^2 value greater than some specified value is equal to the area under the χ^2-distribution curve (tabulated in Table A.5) to the right of this value. This is the same as the t-distribution but different than the table for the normal distribution. The end result means that the χ^2 gives us the probabilities with the greater-than sign, namely: $P(\chi^2 > \chi^2_\alpha) = \alpha$

Statistic 6. the ratio of 2 Variances: Sampling Distribution (F-distribution) WMM (p. 232)

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,

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \qquad (19.9)$$

provides a distribution of the 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.

19.4 Estimation

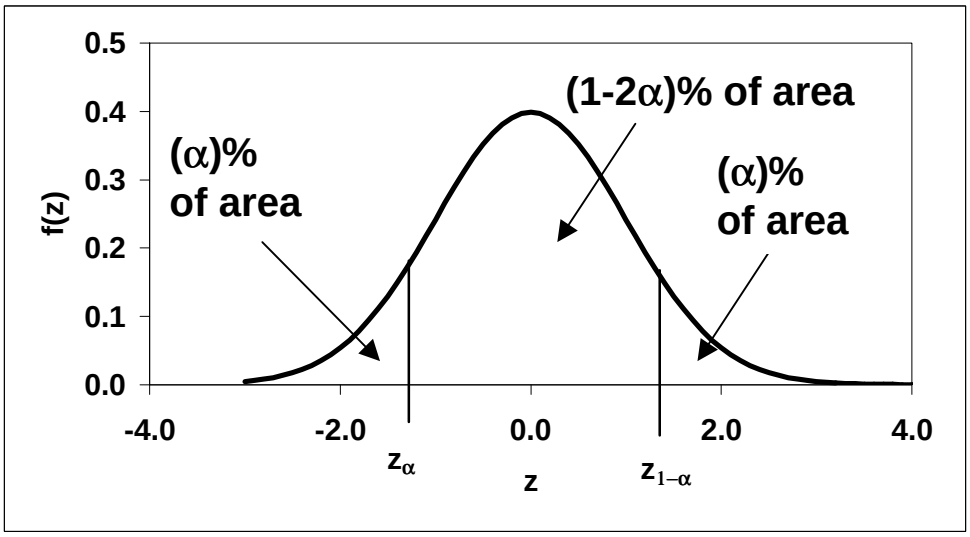
In the previous section we showed what types of distributions describe various statistics of a random sample. In this section, we discuss estimating the population mean and variance from the sample mean and variance. In addition, we introduce confidence intervals to quantify the goodness of these estimates.

19.4.1. Confidence Intervals (p. 240)

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.

In graphical terms:



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

Statistic 1. mean, σ known: confidence interval (p. 243)

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

$$P(\overline{X} + z_\alpha \frac{\sigma}{\sqrt{n}} < \mu < \overline{X} - z_\alpha \frac{\sigma}{\sqrt{n}}) = 1-2\alpha \qquad (19.11)$$

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

Example: 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 the Table A.3, of standard normal PDF values. Therefore, by equation (19.11)

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

so the 95% confidence interval for the mean is

$$5.673 < \mu < 6.327$$

Statistic 3. mean, σ unknown: confidence interval (p. 247)

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 s 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 look up in the t-table once. Additionally, the t table requires you to remember that the t-distribution parameter, v, is v=n-1.

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

where

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

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

$$P(\overline{X} - t_{\alpha}\frac{s}{\sqrt{n}} < \mu < \overline{X} + t_{\alpha}\frac{s}{\sqrt{n}}) = 1-2\alpha \qquad (19.13)$$

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

Example: 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 $\alpha, t_{\alpha}, t_{1-\alpha}$.

$1-2\alpha = 0.95$

$\alpha = 0.025$

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

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

The t value came from the Table A.4, of t-distribution values. Therefore, by equation (19.11)

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

so the 95% confidence interval for the mean is

$$5.662 < \mu < 6.338$$

You see that we are a little less confident about the mean when we use the sample variance as the estimate for the population variance.

Statistic 2. difference of means, σ known: confidence interval (p. 253)

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:

$$P\left[\left(\overline{X}_1-\overline{X}_2\right)+z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}<\left(\mu_1-\mu_2\right)<\left(\overline{X}_1-\overline{X}_2\right)-z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}\right]=1-2\alpha \tag{19.14}$$

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

Example: 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 the Table A.3, of standard normal PDF values. Therefore, by equation (19.14)

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

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

so the 95% confidence interval for the mean is

$$-2.909<\left(\mu_1-\mu_2\right)<-1.091$$

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:

$$P\left[\left(\overline{X}_1-\overline{X}_2\right)-t_\alpha\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}<\left(\mu_1-\mu_2\right)<\left(\overline{X}_1-\overline{X}_2\right)+t_\alpha\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}\right]=1-2\alpha \tag{19.15}$$

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

$$\nu = n_1 + n_2 - 2 \text{ if } \sigma_1 = \sigma_2$$

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

Example: 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 $\alpha, t_\alpha, t_{1-\alpha}$.

$$1 - 2\alpha = 0.95$$

$$\alpha = 0.025$$

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

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

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

The t value came from the Table A.4, of the t-PDF values.

$$P\left[(6-8) - 2.03\sqrt{\frac{1}{36} + \frac{3}{16}} < (\mu_1 - \mu_2) < (6-8) + 2.03\sqrt{\frac{1}{36} + \frac{3}{16}}\right] = 1 - 2(0.025)$$

$$P[-2.942 < (\mu_1 - \mu_2) < -1.058] = 0.95$$

so the 95% confidence interval for the mean is

$$-2.942 < (\mu_1 - \mu_2) < -1.058$$

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 (p. 253)

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

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{\sigma^2} \qquad (19.7)$$

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

So

$$P\left[\frac{(n-1)s^2}{\chi^2_{\alpha}} < \sigma^2 < \frac{(n-1)s^2}{\chi^2_{1-\alpha}}\right] = 1-2\alpha \qquad (19.16)$$

Remember, 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: 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 $\alpha, \chi^2_{\alpha}, \chi^2_{1-\alpha}$.

$$1-2\alpha = 0.95$$
$$\alpha = 0.025$$
$$\chi^2_{\alpha} = \chi^2_{0.025} = 27.488$$
$$\chi^2_{1-\alpha} = \chi^2_{0.975} = 6.262$$

where the parameter v = n-1=35

The t value came from the Table A.5, of χ^2-distribution values.

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

$$P\left[0.5457 < \sigma^2 < 2.395\right] = 0.95$$

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

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{S_1^2\sigma_2^2}{S_2^2\sigma_1^2} \qquad (19.9)$$

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 which says $f_{1-\alpha/2}(v_1,v_2) = \frac{1}{f_{\alpha/2}(v_1,v_2)}$.

So

$$P\left[\frac{S_1^2}{S_2^2}\frac{1}{f_\alpha(v_1,v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{S_1^2}{S_2^2} f_\alpha(v_2,v_1)\right] = 1-2\alpha \qquad (19.17)$$

Example: 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 $\alpha, F_\alpha, F_{1-\alpha}$.

$$1-2\alpha = 0.90$$

$$\alpha = 0.05$$

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

$$F_{1-\alpha} = F_{0.95} = \frac{1}{F_{0.05}(v_2 = 15, v_1 = 19)} = \frac{1}{2.23}$$

with $v_1 = n_1 - 1 = 19$ and $v_2 = n_2 - 1 = 15$

The F value came from the Table A.6, of the F-PDF values.

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

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

so the 90% confidence interval for the mean is

$$0.1431 < \frac{\sigma_1^2}{\sigma_2^2} < 0.7433$$

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.