8 Estimation

8.1 Introduction

Recall that a statistic is a random variable that is a function of the random sample. We call a statistic an *estimator* if it is supposed to estimate or approximate some feature of the population. A *point estimator* is an estimator whose value is a single number and which estimates a single paramter for the population. For example, the sample mean is a point estimator for the population mean. The sample variance is a point estimator for the population variance.

A point estimator only gives an approximation to the parameter it is supposed to estimate. An important question is how good it is. In the following section we consider two characterizations of goodness - bias and mean square error.

8.2 Bias and mean square error for point estimators

Suppose $\hat{\theta}$ is an estimator for the population parameter θ. So $\hat{\theta}$ is a RV and so it has a mean. This mean may or may not be equal to θ.

Definition 1. *$\hat{\theta}$ is an unbiased estimator of θ if $E[\hat{\theta}] = \theta$. If they are not equal we say the estimator is biased. The bias is defined to be*

$$B(\hat{\theta}) = E[\hat{\theta}] - \theta \tag{1}$$

The bias is only part of the story. Even if the estimator is unbiased, if the estimator has a large variance it will not usually be close to θ. So we also need to consider the variance of the estimator. But if the bias is large and the variance is small, the estimator will not usually be close to θ. So rather than just looking at the bias or just looking at the variance, we consider a better measure of goodness - how close $\hat{\theta}$ is to θ on average.

Definition 2. *The mean square error (MSE) of a point estimator for θ is*

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] \tag{2}$$

Proposition 1.

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [B(\hat{\theta})]^2 \tag{3}$$

Proof: To make the following computation a little easier to follow, let $c = E[\hat{\theta}]$. Note that c is a contant, i.e., non-random.

$$\begin{aligned} MSE(\hat{\theta}) &= E[(\hat{\theta}-\theta)^2] \\ &= E[((\hat{\theta}-c)+(c-\theta))^2] \\ &= E[(\hat{\theta}-c)^2+2(\hat{\theta}-c)(c-\theta)+(c-\theta)^2] \\ &= E[(\hat{\theta}-c)^2]+2(c-\theta)E[\hat{\theta}-c]+E[(c-\theta)^2] \\ &= E[(\hat{\theta}-c)^2]+(c-\theta)^2 \\ &= Var(\hat{\theta})+[B(\hat{\theta})]^2 \end{aligned}$$

Definition 3. *The standard error of a point estimator is the standard deviation of the estimator.*

8.3 Some unbiased point estimators

An important problem in statistics is how do you find an estimator for a population parameter. We will study this in chapter 9. Here we look at some intuitively obvious estimators.

The mean of one population: Suppose we have a single population with mean μ. We draw a random sample Y_1 $Y_2, \cdots, Y_n$. The sample mean,

$$\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{4}$$

is a natural estimator for μ. Since $E[\overline{Y}] = \mu$, it is an unbiased estimator. The variance of this estimator is σ^2/n where σ^2 is the population variance. (The book says the variance of the sampling distribution of the estimator is σ^2/n.) Note that the MSE is also σ^2/n since the bias is zero.

The difference of the means of two populations: Suppose we have two populations. Population 1 has mean μ_1 and variance σ_1^2. Population 2 has mean μ_2 and variance σ_2^2. We want to estimate $\mu_1 - \mu_2$. We draw a random sample of size n_1 from population 1 and let $\overline{Y_1}$ denote the sample mean for this sample. We draw an independent random sample of size n_2 from population 2 and let $\overline{Y_2}$ denote the sample mean for this sample. (This is slightly dangerous notation since we have also used a subscript on $\overline{Y}$ to

indicate the sample size. Here the subscript just indicates which population.) Then $\overline{Y_1} - \overline{Y_2}$ is a point estimator for $\mu_1 - \mu_2$. You studied this estimator in homework 2. The mean is $\mu_1 - \mu_2$, so it is an unbiased estimator. The variance is

$$Var(\overline{Y_1} - \overline{Y_2}) = Var(\overline{Y_1}) + (-1)^2 Var(\overline{Y_2}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \tag{5}$$

Since the bias is zero, the MSE is equal to this variance.

End of lecture on Thurs, 1/25

Proportion for one population: Suppose that the experiment only has two outcomes which we will call success and failure. We let p be the probability of success. In terms of the population, we are considering a population in which each member of the population has one of two properties which we call success or failure. An example: the population is all registered voters in the US and we look at whether they voted (success) or not (failure) in the most recent election. In this case p is the fraction of registed voters who voted in the most recent election. Another example: laboratory mice are given a vaccine for a virus and then exposed to the virus. Success is that the mouse gets the virus, failure that the mouse does not. So p is the probability that a vaccinated mouse gets the virus.

The parameter p is the proportion of the population that has success. So we refer to p as the population proportion. We let Y be 1 when we have success and 0 for failure. Note that the mean of Y is just p. Now we take a random sample $Y_1, \cdots, Y_n$ and compute the sample mean in the usual way. Note that $\sum_{i=1}^n Y_i$ is the number of successes in our sample. So $\overline{Y}$ is the proportion of successes in the sample. It is a point estimator for $\mu = p$, the population proportion. So we denote it by $\hat{p}$. The expected value of $\hat{p}$ is $\mu = p$. So the sample proportion $\hat{p}$ is an unbiased estimator of p. The variance of Y is just $p(1-p)$ and so the variance of $\hat{p}$ is $p(1-p)/n$.

Difference of the proportions for two populations: Now suppose we have two populations and the experiments for the populations only have two outcomes. We let p_1 be the proportion for population 1 and p_2 the proportion for population 2. We want to estimate $p_1 - p_2$. We draw a random sample from each population and assume that the two samples are independent of each other. We let $\hat{p_1}$ and $\hat{p_2}$ be the sample proportions for the two random

samples. A natural estimator for $p_1 - p_2$ is $\hat{p_1} - \hat{p_2}$. The mean is $p_1 - p_2$. So it is an unbiased estimator. Since $\hat{p_1}$ and $\hat{p_2}$ are independent RV's, the variance of the estimator is

$$Var(\hat{p_1} - \hat{p_2}) = Var(\hat{p_1}) + Var(\hat{p_2}) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2} \tag{6}$$

An example of the above is the following. We divide our lab mice into a two groups (or populations). Group 1 gets the vaccine. Group 2 (the control group does not. Both groups are exposed to the virus. For both groups success is getting the virus. So p_1 is the probability a vaccinated mouse gets the virus, p_2 the probability an unvaccinated mouse gets it. The parameter difference $p_1 - p_2$ is a measure of the effectiveness of our vaccine.

Finally we consider the variance for a single population. Recall that the sample variance is defined by

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{7}$$

The following proposition says that S^2 is an unbiased estimator for σ^2.

Proposition 2. *Let S^2 be the usual sample variance:*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{8}$$

Then

$$E[S^2] = \sigma^2 \tag{9}$$

So S^2 is an unbiased estimator for the population variance. The variance of S^2 is

$$var(S^2) = \frac{\mu_4}{n} - \frac{\sigma^4(n-3)}{n(n-1)} \tag{10}$$

where $\mu_4 = E[(Y-\mu)^4]$.

Proof:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n} Y_i^2 - n(\overline{Y})^2 \tag{11}$$

$$\tag{12}$$

We use the equation

$$Var(Y) = E[Y^2] - (E[Y])^2 \tag{13}$$

So $E[Y_i^2] = Var(Y_i) + (E[Y_i])^2 = \sigma^2 + \mu^2$. And

$$E[(\overline{Y})^2] = Var(\overline{Y}) + (E[\overline{Y}])^2 = \frac{\sigma^2}{n} + \mu^2 \tag{14}$$

So

$$\begin{aligned} E[\sum_{i=1}^{n}(Y_i - \overline{Y})^2] &= \sum_{i=1}^{n} E[Y_i^2] - nE[(\overline{Y})^2] \\ &= n(\sigma^2 + \mu^2) - n(\frac{\sigma^2}{n} + \mu^2) \\ &= (n-1)\sigma^2 \end{aligned}$$

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

The formula for the variance of S^2 come from a rather long computation which we do not include. □

The estimator in homework problem 2 for the variance was

$$V = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - (\overline{Y})^2 \tag{15}$$

You should have found in the homework that

$$E[V] = \frac{n-1}{n}\sigma^2 \tag{16}$$

So V is a biased estimator of σ^2. Note that the bias goes to zero as $n \to \infty$.

Remark: S^2 is an unbiased estimator for σ^2. Define S to be $\sqrt{S^2}$. This is a natural estimator for σ. However, $E[S]$ is not equal to $(E[S^2])^{1/2} = \sigma$. So S is a biased estimator for σ.

8.4 The goodness of a point estimator

Throughout this section $\hat{\theta}$ is an estimator for the population parameter θ.

Definition 4. *The error of estimation ϵ is defined to be*

$$\epsilon = |\hat{\theta} - \theta| \tag{17}$$

Note that the error of estimation is a random variable. Suppose that our estimator is unbiased. Recall that the standard error is just another name for the standard deviation of the estimator. The main point of this section is to see what we can say about the probability that the error of estimation is less than two times the standard error. We start with an upper bound that always holds.

Theorem 1. (Chebyshev's theorem) *For any RV Y with finite variance σ^2 and mean μ,*

$$P(|Y - \mu| \geq \delta) \leq \frac{1}{\delta^2}\sigma^2 \tag{18}$$

We will use this with $\delta = 2\sigma$. Then it says

$$P(|Y - \mu| \geq 2\sigma) \leq \frac{1}{4} \tag{19}$$

This is the same as

$$P(|Y - \mu| < 2\sigma) \geq \frac{3}{4} \tag{20}$$

So for any random variable the probablity the RV is within two standard deviations of the mean is at least 0.75. For most distributions, and especially those that occur as the distribution of an estimator, this probability is much closer to 1. A good rule of thumb is that this probability is around 0.95. For a normal distribution this probability is $P(|Z| \geq 2) \approx 0.9544$ (where Z is standard normal). For a uniform distribution this probabiilty is 1.

An example of a distribution for which the probability is essentially 0.75 is the following. Let Y be a discrete RV which only takes on the three values $-1, 0, 1$. We take

$$P(Y = 0) = 1 - p, P(Y = 1) = P(Y = -1) = p/2 \tag{21}$$

The mean is 0. The second moment is $E[Y^2] = p$. So $\sigma = \sqrt{p}$. If $2\sqrt{p} < 1$, then the probability that Y is more than 2σ from the mean is just the probability that $Y = \pm 1$ which is p. The condition $2\sqrt{p} < 1$ is the same as $p < 1/4$. So if we take p just slightly smaller than $1/4$, then the probability that Y is more than 2σ from the mean is almost $1/4$. This example shows that although the Chebyshev bound is usually far from the truth, there are distributions for which you cannot do any better.

Returning to estimators, we apply the above bound with $Y = \hat{\theta}$ and $\delta = 2\sigma_\theta$. Note that $\sigma_{\hat{\theta}}$ is the standard error. Since the estimator is unbiased, the mean of $\hat{\theta}$ is θ. So we get

$$P(|\hat{\theta} - \theta| < 2\sigma_{\hat{\theta}}) \geq \frac{3}{4} \tag{22}$$

Thus for any unbiased estimator, with probability at least 0.75, the error of estimation is no more than twice the standard error. And for most estimators the probability is closer to 1, often around 0.95.

Example: Consider the mice and vaccine example from previous section. 100 mice are given vaccine and exposed to a virus. 15 of then get the virus. 50 control mice are not given the vaccine and exposed to the virus. 42 of them get the virus. Estimate $p_2 - p_1$ and give a two standard error bound on the estimate.

We have $\hat{p_1} = 15/100 = 0.15$ and $\hat{p_2} = 42/50 = 0.84$. So our estimate for $p_2 - p_1$ is $0.84 - 0.15 = 0.69$. The standard error is given by $\sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2}$. We don't know p_1 and p_2, so we have to approximate them with $\hat{p_1}$ and $\hat{p_2}$. So the standard error is approximately

$$\sqrt{0.15(1-0.15)/100 + 0.84(1-0.84)/50} = 0.0629 \tag{23}$$

So 2 times the standard error is 0.126. So the true value of $p_2 - p_1$ will be within 0.126 of 0.69 with probability of at least 0.75. The sample sizes are pretty large, so by the CLT, our estimator $\hat{p_2} - \hat{p_1}$ will be approximately normal. So the probability our estimate is within 0.126 of 0.69 should actually be about 0.95.

End of lecture on Tues, 1/30

So far we have only considered unbiased estimators for which the error of estimation is $|\hat{\theta} - \theta| = |\hat{\theta} - E[\hat{\theta}]|$. What if the estimator is biased? Can we get a similar statement? We start with another version of Chebyshev.

Theorem 2. (Chebyshev's theorem - another version) *For any RV Y with finite variance σ^2 and any constant c*

$$P(|Y - c| \geq \delta) \leq \frac{1}{\delta^2} E[(Y - c)^2] \tag{24}$$

Taking $c = \mu$ gives the version we saw before. Now we apply this with $Y = \hat{\theta}$, $c = \theta$, and $\delta = 2\sqrt{MSE}$. Recall that the MSE is $E[(\hat{\theta} - \theta)^2]$. So the above becomes

$$P(|\hat{\theta} - \theta| \geq 2\sqrt{MSE}) \leq \frac{1}{4} \tag{25}$$

Note that if the estimator is unbiased, then $\sqrt{MSE}$ is the standard error and this becomes the same bound we saw before. As before this probability is typically much closer to 0 than 1/4. In many case it is about 0.05. So for a biased estimator, the error is estimation is no more than $2\sqrt{MSE}$ with probabiity at least 75% and more typically 95%.

8.5 Confidence intervals

A confidence interval for θ is an estimator for θ which is not a point estimator. Instead the estimator gives an interval. We denote this interval by $[\hat{\theta}_L, \hat{\theta}_U]$.

Definition 5. *The confidence coefficient is*

$$P(\theta \in [\hat{\theta}_L, \hat{\theta}_U]) \tag{26}$$

The confidence coefficient is a measure of how realiable our confidence intervals is. The closer the confidence coefficient is to 1, the more reliable the confidence interval. Note that if we make the confidence interval larger, then the confidence coefficient will increase. But to say μ belongs to a large confidence interval is less desirable that being able to say it belongs to a small confidence interval. So there is a trade-off. Typical choices for the confidence coefficient are 0.9, 0.95, 0.98 or maybe even 0.99. Confidence coefficients are often written in the form $1 - \alpha$. So α is typically $0.1, 0.05, 0.02, 0.01$.

We can also consider one-sided confidence intervals, i.e., $(-infty, \hat{\theta}_U]$ or $[\hat{\theta}_L, \infty)$. For these the confidence coefficient would be $P(\theta \leq \hat{\theta}_U)$ or $P(\theta \geq \hat{\theta}_L)$, respectively.

Given a particular confidence coefficient $1 - \alpha$ we want to find the confidence interval, i.e., $\hat{\theta}_L$ and $\hat{\theta}_U$. Note that $\hat{\theta}_L$ and $\hat{\theta}_U$ must be functions of the

random sample and constants like the sample size, but they cannot depend on unknown population parameters like θ. So we want to find estimators $\hat{\theta}_L$ and $\hat{\theta}_U$ so that

$$P(\theta \in [\hat{\theta}_L, \hat{\theta}_U]) = 1 - \alpha \tag{27}$$

This is the same as

$$P(\theta < \hat{\theta}_L) + P(\theta > \hat{\theta}_U) = \alpha \tag{28}$$

There is freedom here as to how much or α goes into each of the two probabilities. The simplest thing to do is require

$$P(\theta < \hat{\theta}_L) = \alpha/2, \quad P(\theta > \hat{\theta}_U) = \alpha/2 \tag{29}$$

The difficulty here is that the distribution of the estimators $\hat{\theta}_L$ and $\hat{\theta}_U$ will typically depend on θ which we do not know. To deal with this difficulty we introduce the idea of a *pivotal quantity.*

Definition 6. *A pivotal quantity (for the parameter θ) is a function of the random sample with two properties*

- *The pivotal quantity depends on the random sample and the unknown parameter θ, but it does not depend on any other unknown parameters.*
- *The distribution of the pivotal quantity does not depend on θ.*

Given a pivotal quantity U, the strategy for using it to get a confidence interval is the following. Since the distribution of U does not depend on θ, we can find numbers a and b so that $P(a \leq U \leq b) = 1 - \alpha$. Since U depends on the random sample and θ, the inequality $a \leq U \leq b$ can be transformed into a statement that θ belows to some interval whose endpoints are functions of the random sample. We illustrate with two examples.

Example: We first do an unrealistic example to illustrate the idea. Suppose that the population has an exponential distribution. So Y has an exponential distribution. To be consistent with the notation in this section we let θ be the parameter for the exponential. So the mean of Y is θ. We do not know θ and want a confidence interval for it. We take a random sample of size 1. So the random sample is just Y_1. We claim that Y_1/θ is a pivotal quantity. It clearly meets the first condition. We need to show that its distribution does not depend on θ. SHOW THIS.

Now we use this pivotal quantity to find a confidence interval for θ. We take $\alpha = 0.1$. We start by finding an interval $[a, b]$ so that the pivotal quantity is in $[a, b]$ with probability 0.9. Let $U = Y_1/\theta$ denote the pivotal quantity. So we want

$$P(a \leq U \leq b) = 0.9 \tag{30}$$

There are many a, b that satisfy this. We will ask that

$$P(U < a) = 0.05, \quad P(U > b) = 0.05 \tag{31}$$

The distribution of U is just an exponential with mean 1. So $a = qexp(0.05, 1) = 0.05129$. And $b = qexp(0.95, 1) = 2.9957$. Now we use the fact that the pivotal quantity depends on θ but no other unknown parameter to turn the statement that $a \leq U \leq b$ into a statement about θ:

$$\begin{aligned} \{a \leq U \leq b\} &= \{a \leq Y_1/\theta \leq b\} = \{1/b \leq \theta/Y_1 \leq 1/a\} & (32) \\ &= \{Y_1/b \leq \theta \leq Y_1/a\} = \{\theta \in [Y_1/b, Y_1/a]\} & (33) \end{aligned}$$

So

$$P(\theta \in [Y_1/b, Y_1/a]) = 0.9 \tag{34}$$

We have found a confidence interval for θ with confidence coefficient 0.9. It is $[Y_1/b, Y_1/a] = [0.3338Y_1, 19.50Y_1]$. Note that this confidence interval is huge. But that is to be expected if we use a sample of size 1.

Example(continued): We continue to consider a Y that has an exponential distribution with mean θ. But now we take a sample of size n. In a homework problem we worked out the pdf of $\overline{Y}_n$ in this case. It has a gamma distribution with $\alpha_{\overline{Y}} = n$ and $\beta_{\overline{Y}} = \theta/n$. We claim that $U = \overline{Y}_n/\theta$ is a pivotal quantity. It clearly meets the first condition. We need to show that its distribution does not depend on θ. To show this, recall that the mgf of a gamma distribution with parameters α, β is

$$M(t) = \left(\frac{1}{1-\beta t}\right)^{\alpha} \tag{35}$$

SHOW the distribution of U is a gamma distribution with $\alpha = n$ and $\beta = 1/n$. We now look for a and b such that

$$P(U < a) = 0.05, \quad P(U > b) = 0.05 \tag{36}$$

The values of a and b will depend on n. To be concrete we work this out for sample sizes of $n = 10, 100$. Note that in R, the arguments for the gamma routine are a bit confusing. The default arguments are not α, β but rather $\alpha, 1/\beta$. For example, to compute the value of the cdf at 0.05 for $\alpha = 10$, $\beta = 0.1$ we use $qgamma(0.05, 10, 10)$ You can force it to use β instead of $1/\beta$ by $qgamma(0.05, 10, scale = 0.1)$. (scale is another name for the parameter β.)

$$\begin{aligned} a_{10} &= qgamma(0.05, 10, 10) = 0.5425 \\ b_{10} &= qgamma(0.95, 10, 10) = 1.5705 \\ a_{100} &= qgamma(0.05, 100, 100) = 0.8414 \\ b_{100} &= qgamma(0.95, 100, 100) = 1.170 \end{aligned}$$

So the 90% confidence intervals are $[0.6367\overline{Y_{10}}, 1.843\overline{Y_{10}}]$ for $n = 10$ and $[0.8547\overline{Y_{100}}, 1.1885\overline{Y_{100}}]$ for $n = 100$. Note that as the sample size gets larger, the confidence interval gets smaller.

8.6 Large sample confidence intervals

In this section we consider confidence intervals corresponding to the point estimators we studied earlier : $\hat{\mu}, \hat{\mu_1} - \hat{\mu_2}, \hat{p}, \hat{p_1} - \hat{p_2}$ under the assumption that the sample size(s) are large. If the sample size is large, than the distribution of all of these estimators is approximately normal. Recall that all four of these estimators are unbiased. Let $\hat{\theta}$ be one of these four estimators and θ the corresponding population parameter. So $E[\hat{\theta}] = \theta$. If we define

$$Z = \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \tag{37}$$

then the distribution of Z will be approximately the standard normal. This is one of the criteria for a pivotal quantity. Z does not quite meet the first criteria since it contains the unknown parameter $\sigma_{\hat{\theta}}$ in addition to the unknown parameter θ. But for large samples we can approximate $\sigma_{\hat{\theta}}$ by a function of the random sample that does not involve any unknown parameters. We return to this point later.

Now suppose we want a confidence interval with confidence coefficient $1 - \alpha$. We start by looking for a and b so that

$$P(Z < a) = \alpha/2, \quad P(Z > b) = \alpha/2 \tag{38}$$

Since the distrbution of Z is symmetric, a will just be $-b$. The usual notation is to write b as z_c. Here are some commonly used values of α and the corresponding z_c:

$$\alpha = 0.1 \Rightarrow z_c = 1.645..., \tag{39}$$
$$\alpha = 0.05 \Rightarrow z_c = 1.960..., \tag{40}$$
$$\alpha = 0.01 \Rightarrow z_c = 2.576..., \tag{41}$$
$$\tag{42}$$

So we now have $P(-z_c \leq Z \leq z_c) = 1 - \alpha$. Using the definition of Z this becomes

$$P(-z_c \leq \frac{\hat{\theta} - \theta}{\sigma_{\hat{\theta}}} \leq z_c) = 1 - \alpha \tag{43}$$

which we can rewrite as

$$P(\theta \in [\hat{\theta} - z_c\sigma_{\hat{\theta}}, \hat{\theta} + z_c\sigma_{\hat{\theta}}) = 1 - \alpha \tag{44}$$

So the confidence interval is

$$[\hat{\theta} - z_c\sigma_{\hat{\theta}}, \hat{\theta} + z_c\sigma_{\hat{\theta}}] \tag{45}$$

This result applies to each of the four estimators $\hat{\mu}, \hat{\mu_1} - \hat{\mu_2}, \hat{p}, \hat{p_1} - \hat{p_2}$.

We now return to the issue of the dependence of $\sigma_{\hat{\theta}}$ on unknown parameters. First suppose $\theta = \mu$ and $\hat{\theta} = \overline{Y}$. For this case $\sigma_{\hat{\theta}}$ is $\sigma/\sqrt{n}$, and σ is unknown. We approximate σ by the sample standard deviation S. So our confidence interval is

$$[\overline{Y} - z_c S/\sqrt{n}, \overline{Y} + z_c S/\sqrt{n}] \tag{46}$$

End of lecture on Thurs, 2/1

If $\theta = \mu_1 - \mu_2$ and $\hat{\theta} = \overline{Y_1} - \overline{Y_2}$, then we make the approximation

$$\sigma_{\hat{\theta}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \approx \sqrt{S_1^2/n_1 + S_2^2/n_2} \tag{47}$$

where S_1^2 is the sample variance of the sample from population 1 and S_1^2 is the sample variance of the sample from population 2. So our confidence interval is

$$[\overline{Y_1} - \overline{Y_2} - z_c\sqrt{S_1^2/n_1 + S_2^2/n_2}, \overline{Y_1} - \overline{Y_2} + z_c\sqrt{S_1^2/n_1 + S_2^2/n_2}] \tag{48}$$

If $\theta = p$ and $\hat{\theta} = \hat{p}$, then we make the approxmation

$$\sigma_{\hat{p}} = \sqrt{p(1-p)/n} \approx \sqrt{\hat{p}(1-\hat{p})/n} \tag{49}$$

So the confidence interval is

$$\hat{p} \pm z_c\sqrt{\hat{p}(1-\hat{p})/n} \tag{50}$$

Finally, for $\theta = p_1 - p_2$ and $\hat{\theta} = \hat{p_1} - \hat{p_2}$, we make the approxmation

$$\sigma_{\hat{p}} = \sqrt{p_1(1-p_1)/n_1 + p_2(1-p_2)/n_2} \approx \sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2} \tag{51}$$

So the confidence interval is

$$\hat{p_1} - \hat{p_2} \pm z_c\sqrt{\hat{p_1}(1-\hat{p_1})/n_1 + \hat{p_2}(1-\hat{p_2})/n_2}] \tag{52}$$

Example: We want to compare the average height of an adult female in Canada and the US. We randomly choose 1000 adult females in Canada (pop 1) and 1000 adult females in the US (pop 2). We find

$$\overline{Y_1} = 161.0cm, \quad S_1 = 12.2cm \tag{53}$$
$$\overline{Y_2} = 164.1cm, \quad S_2 = 13.1cm \tag{54}$$
$$\tag{55}$$

Find a 95% confidence interval for $\mu_1 - \mu_2$.

$$\sigma^2_{\overline{Y_1}-\overline{Y_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \approx \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} = \frac{(12.2)^2}{1000} + \frac{(13.1)^2}{1000} = 0.320 \tag{56}$$

So $\sigma_{\overline{Y_1}-\overline{Y_2}} = 0.566$. For 95% confidence interval $z_c = 1.96$. So confidence interval is

$$(161.0 - 164.1) \pm 1.96 * 0.566 = -3.1 \pm 1.109 \tag{57}$$

Example: (from the book) We have two brands of refrigerators. Both have a one year warranty. For brand A we find that in a sample of 50, 12 fail during the warranty period. For brand B we find that in a sample of 60, 12 fail during the warranty period. We define p_A, p_B to be the probability of a failure. Estimate $p_A - p_B$ with a confidence interval with confidence coefficient of 98%.

$$\hat{p_A} = 12/50 = 0.24, \hat{p_B} = 12/60 = 0.20 \tag{58}$$

Critical z is 2.33. Confidence interval is

$$\begin{aligned} (0.24 - 0.20) &\pm 2.33\sqrt{\frac{(0.24)(0.76)}{50} + \frac{(0.20)(0.80)}{60}} && (59) \\ &= 0.04 \pm 0.1851 = [-0.1451, 0.2251] && (60) \end{aligned}$$

8.7 Choosing the sample size

If we have a sample of size n and we want a certain confidence coefficient $1 - \alpha$, then we have seen how to find the confidence interval $[\hat{\theta}_L, \hat{\theta}_U]$ for several different estimates. Here we consider a different question. Suppose we have not drawn the sample yet and we want a given confidence coeffecient and we want the confidence interval to be a certain size. Note that $\hat{\theta}_L$ and $\hat{\theta}_U$ depend on the sample size, and so we can use the given width for the confidence interval to solve for the sample size.

Example We continue the example comparing the average heights of adult females in the US and Canada. With sample sizes of 1000 we found a confidence interval of -3.1 ± 1.109. So the width was 2.22. Suppose we want the width of a 95% confidence interval to be 1.0 and we are going to use samples of the same size. What should n be?

Answer: We had

$$S_1 = 12.2cm, \qquad S_2 = 13.1cm \tag{61}$$

$$\tag{62}$$

We continue to approximate

$$\sigma^2_{\overline{Y_1} - \overline{Y_2}} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \approx \frac{S_1^2}{n} + \frac{S_2^2}{n} = \frac{(12.2)^2}{n} + \frac{(13.1)^2}{n} = \frac{320.45}{n} \tag{63}$$

So $\sigma_{\overline{Y_1}-\overline{Y_2}} = 17.9/\sqrt{n}$. For 95% confidence interval $z_c = 1.96$. The width of the confidence interval will be $2z_c\sigma_{\overline{Y_1}-\overline{Y_2}} = 70.17/\sqrt{n}$. So $n = (70.17)^2 = 4925$.

Example A marketing firm wants to test interest in a new product. They will ask the people sampled one question - would they buy the product or not. They want to estimate the fraction p of people who say yes with a 90% confidence interval and want the estimate to be within 2% of the true value, i.e., they want the width of the confidence interval to be 4%.

Answer: We have no idea what p really is, so how do we find $\sigma_{\hat{p}}$? It is given by $\sqrt{p(1-p)}/\sqrt{n}$. The maximum of $p(1-p)$ over $0 \leq p \leq 1$ occurs when $p = 1/2$. So we have $\sigma_{\hat{p}} \leq 1/(2\sqrt{n})$. For a 90% confidence interval, $z_c = 1.645$. We want $z_c/(2\sqrt{n}) = 0.02$. This leads to $n = 1692$.

8.8 Small sample confidence intervals

In the section on large sample confidence intervals, we used the fact that the sample was large in two ways. First, a large sample size implies that $\overline{Y}$ will be approximately normally distributed. Second, a large sample size allowed us to approximate the unknown population variance by the sample variance.

We now assume the sample size is not large. We assume the population is normal (or at least close to normal). We review the definition of the t-distribution. Let Z and W be independent RV's and suppose Z has a standard normal distribution and W has a χ^2 distribution with ν degrees of freedom. Then

$$T = \frac{Z}{\sqrt{W/\nu}} \tag{64}$$

has a t-distribution with ν degrees of freedom. If the population is normal, then $\overline{Y}$ is normal with mean μ and variance σ^2/n. And $(n-1)S^2/\sigma^2$ has a χ^2 distribution with $n-1$ degrees of freedom. So

$$T = \frac{\overline{Y} - \mu}{\sqrt{S^2/n}} \tag{65}$$

has a t-distribution with $n-1$ degrees of freedom.

The statistic T depends on the random sample through $\overline{Y}$ and S^2. It depends on only one unknown parameter, namely μ. And the distribution

of T does not depend on μ. So T is a pivotal quantity for μ. So we can use it to contruct a confidence interval for μ with confidence coefficient $1-\alpha$. Recall that the t-distribution is symmetric. So when we seek a and b such that $P(a \leq T \leq b)$, it is natural to take $a = -b$. Let $t_{\alpha/2}$ be defined by

$$P(T > t_{\alpha/2}) = \alpha/2 \tag{66}$$

Note that $t_{\alpha/2}$ depends on the number of degree of freedom, $n-1$. So we have

$$P(-t_{\alpha/2} \leq T \leq t_{\alpha/2}) = \alpha \tag{67}$$

The next step in finding the confidence interval is to rearrange $-t_{\alpha/2} \leq T \leq t_{\alpha/2}$ so that it gives a confidence interval for μ. We find that $-t_{\alpha/2} \leq T \leq t_{\alpha/2}$ is equivalent to

$$\mu \in [\overline{Y} - t_{\alpha/2}S/\sqrt{n}, \overline{Y} + t_{\alpha/2}S/\sqrt{n}] \tag{68}$$

So the confidence interval with confidence coefficient $1-\alpha$ is

$$\overline{Y} \pm t_{\alpha/2}S/\sqrt{n} \tag{69}$$

End of lecture on Tues, 2/6

Example: A random sample of 16 Americans found that the amount of beef they ate (in lbs) in the past year was,

$$118, 115, 125, 110, 112, 130, 117, 112, 115, 120, 113, 118, 119, 122, 123, 126$$

You can enter this in R with

$$x = c(118, 115...126) \tag{70}$$

Then

$$mean(x) = 118.4, \quad sd(x) = 5.656, \quad qt(0.025, 15) = -2.131 \tag{71}$$

So we get a confidence interval of 118.4 ± 3.106. Note that if we had pretended 16 was a big sample we would have used 1.96 in place of 2.131 and gotten 118.4 ± 2.772.

We now consider a confidence interval for the difference of two population means $\mu_1 - \mu_2$ when the sample sizes are not large. We assume that each of the two populations are normal (or close to normal) and that the samples drawn from the populations are independent or each other. We know that $\overline{Y_1}$ is independent of S_1^2 and that $\overline{Y_2}$ is independent of S_2^2. We also know that $(n_1 - 1)S_1^2/\sigma_1^2$ and $(n_2 - 1)S_2^2/\sigma_2^2$ have χ^2 distributions. We want to replace

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \to \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \tag{72}$$

But ...

If we assume that the two populations have the same variance, then we can make progress. So we assume $\sigma_1^2 = \sigma_2^2$. We denote this common variance by σ^2. We estimate this variance by "pooling our two samples." Define

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 - 1 + n_2 - 1} \tag{73}$$

Note that $(n_1 + n_2 - 2)S_p^2/\sigma^2$ is the sum of two independent χ^2 distributions with $n_1 - 1$ and $n_2 - 1$ d.f. So it has a χ^2 distribution with $n_1 + n_1 - 2$ d.f. Also, it is independent of $\overline{Y_1} - \overline{Y_2}$. After a bit of algebra this implies

$$T = \frac{\overline{Y_1} - \overline{Y_2} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \tag{74}$$

has a t-distribution with $n_1 + n_2 - 2$ d.f. It is a pivotal quantity for $\mu_1 - \mu_2$. So we can use it to find a confidence interval for $\mu_1 - \mu_2$. We define $t_{\alpha/2}$ as before. The event $-t_{\alpha/2} \leq T \leq t_{\alpha/2}$ leads to the confidence interval

$$(\overline{Y_1} - \overline{Y_2}) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} \tag{75}$$

Example: A drug is supposed to lower blood pressure. 10 subjects are divided into two groups. 6 subjects (group 1) are given the drug over some period of time and 4 subjects (group 2) are not. At the end of the trial their systolic blood pressure is measured.

$$\overline{Y_1} = 117.5, \quad S_1 = 9.7, \tag{76}$$

$$\overline{Y_2} = 126.8, \quad S_1 = 12.0 \tag{77}$$

Find a 95% confidence interval for $\mu_1 - \mu_2$. Does this study provide evidence that the drug lowers blood pressure?

Describe what population 1 and 2 are.

We assume that the variances of the the two populations are the same. The pooled estimator for this common variance is

$$S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1 - 1 + n_2 - 1} \tag{78}$$

$$= \frac{(6-1)(9.7)^2 + (4-1)(12.0)^2}{6-1+4-1} = 112.8 \tag{79}$$

So $S_p = 10.6$. The number of degrees of freedom is $6+4-2 = 8$. The critical t_c is given by

$$t_c = -qt(0.025, 8) = 2.306 \tag{80}$$

The confidence interval works out to -9.3 ± 15.77. We cannot conclude from this that $\mu_1 - mu_2 < 0$. So this study does not provide evidence that the drug lowers blood pressure. (This does not mean that the study shows it does not lower blood pressure.)

8.9 Confidence interval for σ^2

In this section we consider a confidence interval for the population variance σ^2. We have already seen a point estimator for this parameter, namely, the sample variance

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{81}$$

We assume that the population is normal. Recall that under this assumption

$$U = \frac{(n-1)S^2}{\sigma^2} \tag{82}$$

has a χ^2 distribution with $n-1$ degrees of freedom. Note also that U only involves one unknown parameter, σ^2. So U is a pivotal quantity. So to find a confidence interval for σ^2 we first seek numbers χ_L^2 and χ_U^2 such that

$$P(\chi_L^2 \leq U \leq \chi_U^2) = 1 - \alpha \tag{83}$$

As we have encountered before, there are many solutions of this equation. One natural way to pick out one would be to find the solution that minimizes $\chi^2_U - \chi^2_L$. This will lead to the smallest possible confidence interval. But there is no explicit way to compute this solution. We could find it by trial and error.

We take a simpler approach and choose χ^2_u and χ^2_L so that

$$P(U < \chi^2_L) = \alpha/2, \quad P(U > \chi^2_U) = \alpha/2 \tag{84}$$

Using the definition of U, eq. (83) implies

$$P((n-1)S^2/\chi^2_U \leq \sigma^2 \leq (n-1)S^2/\chi^2_U) = 1 - \alpha \tag{85}$$

So our confidence interval for σ^2 is

$$[(n-1)S^2/\chi^2_U, \leq (n-1)S^2/\chi^2_L] \tag{86}$$

Often χ^2_L is written as $\chi^2_{\alpha/2,n-1}$ and χ^2_U is written as $\chi^2_{1-\alpha/2,n-1}$.

Example: A population is approximately normally distributed. I want to estimate the population variance σ^2. I pick a random sample of size 10 and find a sample mean of 102.7 and a sample variance of 4.2. Find a 95% confidence interval for the population variance σ^2. We have $\alpha = 0.05$. So

$$\chi^2_L = qchisq(0.025, 9) = 2.70, \quad \chi^2_U = qchisq(0.975, 9) = 19.0, \tag{87}$$

So the confidence interval is

$$[\frac{9 \times 4.2}{19.0}, \frac{9 \times 4.2}{2.70}] = [1.99, 14.0] \tag{88}$$

We can get a 95% confidence interval for σ by taking square roots $[\sqrt{1.99}, \sqrt{14}] = [1.41, 3.74]$.

End of lecture on Thurs 2/8
