Section 4.2. Confidence Intervals

Note. This section is a continuation of Section 4.2. Sampling and Statistics. We still want to approximate a density function $f(x;\theta)$ of random variable X where θ is unknown. Now, based on a sample $X_1, X_2, \ldots, X_n$, we estimate θ as $\hat{\theta} = \hat{\theta}(X_1, X_2, \ldots, X_n)$. We find confidence intervals for θ. In particular, we explore confidence intervals for the mean of a normal distribution (in Example 4.2.1), for the parameter p for a Bernoulli random variable (in Example 4.2.3), for the difference of two means, and for the difference in two proportions in Bernoulli distributions.

Definition 4.2.1. Let $X_1, X_2, \ldots, X_n$ be a sample on a random variable X where X has probability density function $f(x;\theta)$ where $\theta \in \Omega$. Let $0 < \alpha < 1$ be given. Let $L = L(X_1, X_2, \ldots, X_n)$ and $U = U(X_1, X_2, \ldots, X_n)$ be two statistics. The open interval (L, U) is a $(1-\alpha)100\%$ *confidence interval* for θ if $1 - \alpha = P_\theta(\theta \in (L, U))$. That is, the probability that the interval includes θ is $1-\alpha$. The value of $1-\alpha$ is the *confidence coefficient* or *confidence level* of the interval.

Note. With the sample realized as $x_1, x_2, \ldots, x_n$, then L and U are estimated as l and u. Then the interval (l, u) either includes θ or it does not. We can then think of this as a Bernoulli trial with probability of success (where "success" means $\theta \in (l, u)$) of $1-\alpha$. If M independent $1-\alpha$ level confidence intervals are considered, we expect $(1-\alpha)M$ of the intervals to include θ (since the mean of the binomial distribution is the probability of success times the number of trials, by Note 3.1.A). In this case, we have $(1-\alpha)100\%$ confidence that θ lies in (l, u).

Note/Definition. We may determine confidence intervals in different ways. The *efficiency* of a confidence interval is its expected length. So for (L_1, U_1) and (L_2, U_2) as two confidence intervals for θ that have the same confidence coefficient, $(L{-}1, U_1)$ is more efficient that (L_2, U_2) if $E_\theta(u_1 - l_1) \leq E_\theta(u_1 - l_2)$ for all $\theta \in \Omega$.

Note. We now give some examples of procedures for finding confidence intervals. In this section, we consider a "pivot random variable" which is a function of θ and x and the distribution of the pivot is known. Finding confidence intervals without the use of distributions is considered in Section 4.4. Order Statistics.

Example 4.2.1. In this example, we consider a random sample $X_1, X_2, \ldots, X_n$ from a distribution which is known to be (some) normal distribution $N(\mu, \sigma^2)$. We want a confidence interval for μ. Let $\overline{X}$ and X^2 be the sample mean and sample variance (computed using "$n-1$"). As shown in Example 4.1.3, μ is the maximum likelihood estimator of $\overline{X}$ and $((n-1)/n)S^2$ is the maximum likelihood estimates of σ^2. By Theorem 3.6.1(d) (Student's Theorem), random variable $T = (\overline{X}-\mu)/(S/\sqrt{n})$ has a Student's t-distribution with $n-1$ degrees of freedom. Here, T is the "pivot variable." For given $0 < \alpha < 1$, let $t_{\alpha/2,n-1}$ be the $\alpha/2$ critical point of the t-distribution with $n-1$ degrees of freedom so that $\alpha/2 = P(T > t_{\alpha/2,n-1})$. Since the t-distribution is symmetric about the y-axis then we have

$$\begin{aligned}
1-\alpha &= P(-t_{\alpha/2,n-1} < T < t_{\alpha/2,n-1}) = P\left(-t_{\alpha/2,n-1} < \frac{\overline{X}-\mu}{S/\sqrt{n}} < t_{\alpha/2,n-1}\right) \\
&= P(-t_{\alpha/2,n-1}S/\sqrt{n} < \overline{X}-\mu < t_{\alpha/2,n-1}S/\sqrt{n}) \\
&= P(-\overline{X} - t_{\alpha/2,n-1}S/\sqrt{n} < -\mu < -\overline{X} + t_{\alpha/2,n-1}S/\sqrt{n})
\end{aligned}$$

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

Let $\overline{x}$ and s denote the realized values of $\overline{X}$ and S based on the realized sample. Then a $(1-\alpha)100\%$ confidence interval for μ is

$$(\overline{x} - t_{\alpha/2,n-1}s/\sqrt{n}, \overline{x} + t_{\alpha/2,n-1}x/\sqrt{n}).$$

This is called the $(1-\alpha)100\%$ *t-interval* for μ and $s/\sqrt{n}$ is the *standard error* of $\overline{X}$.

Exercise 4.2.1. Let the observed value of the mean $\overline{X}$ and the sample variance of a ransom sample of size 20 from a distribution that is $N(\mu, \sigma^2)$ be 81.2 and 25.6, respectively. First, the 90%, 95%, and 99% confidence intervals for μ. Not how the lengths of the confidence intervals increase as the confidence increases.

Note. The confidence intervals above require that the samples are normally distributed. However, even if the sampled *population* is not normally distributed, the *samples* are approximately normal distributed by the Central Limit Theorem (Theorem 5.3.1). We restate the Central Limit Theorem here (or state it for the first time if you are going through the chapters in order; in Chapter 5 it is stated in terms of convergence in measure but is stated here simply in terms of limits).

Theorem 4.2.1. Central Limit Theorem.

Let $X_1, X_2, \ldots, X_n$ denote the observations of a random sample from a distribution that has more μ and finite variance σ^2. Then the distribution function of the random variable $W_n = (\overline{X} - \mu)/(\sigma/\sqrt{n})$ converges to Φ, the standard normal distribution $N(0,1)$, as $n \to \infty$.

Example 4.2.2. Consider a random sample $X_1, X_2, \ldots, X_n$ on a random variable X with mean μ and variance σ^2 where the distribution of X is not normal or is unknown. Since the distribution of $Z_n = (\overline{X} - \mu)/(S/\sqrt{n})$ is approximately $N(0,1)$ (with the approximation "good" for n "large," as described above), then

$$1 - \alpha \approx P\left(-x_{\alpha/2} < \frac{\overline{X} - \mu}{S/\sqrt{n}} < z_{\alpha/2}\right) = P\left(\overline{X} - z_{\alpha/2}S/\sqrt{n} < \mu < \overline{X} + z_{\alpha/2}S/\sqrt{n}\right).$$

This is the *large sample confidence interval* for μ.

Note. For the same α, the confidence intervals based on the t-distribution and $t_{\alpha/2,n-1}$ are larger than those based on the standard normal distribution and $z_{\alpha/2}$. So the t0interval is more conservative than the large sample confidence interval.

Example 4.2.3. Consider a random sample $X_1, X_2, \ldots, X_n$ on a Bernoulli random variable X with probability of success p (where $X = 1$ is a success and $X = 0$ is a failure). The sample average is $\overline{X} = \frac{1}{n}\sum_{i=1}^n X_i$ and we take this as $\hat{p} = \overline{X}$. The sample variance is $\text{Var}(\hat{p}) = \sigma^2/n = p(1-p)/n$. So by the Central Limit Theorem we have that random variable $Z = (\hat{p} - p)/\sqrt{p(1-p)/n}$ is approximately $N(0,1)$ for n "large." Estimating σ^2 with $S^2 = \hat{p}(1-\hat{p})$, we have the $(1-\alpha)100\%$ confidence interval for p as

$$\left(\hat{r} - z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}, \hat{r} + z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}\right).$$

here, $\sqrt{\hat{p}(1-\hat{p})/n}$ is called the *standard error* of $\hat{p}$.

Exercise 4.2.A. (From William Navadi's *Statistics for Engineers and Scientists*, 3rd Edition, McGraw-Hill [2011]). In a simple random sample of 70 automobiles registered in a certain state, 28 or them were found to have emission levels that exceed a state standard. Find 95% and 98% confidence intervals for the proportion of automobiles in the state whose emission levels exceed the standard.

Note. We now consider the comparison of the means of two random variables X and Y. We do so using samples and confidence intervals for the difference of the means. Let μ_1 and μ_2 be the means of X and Y, respectively, and define $\Delta = \mu_1 - \mu_2$. Assume the variances of X and Y are finite and let their variances $\sigma_1^2 = \text{Var}(X)$ and $\sigma_2^2 = \text{Var}(Y)$, respectively. Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample on X and let $Y_1, Y_2, \ldots, Y_{n_2}$ be a random sample on Y. Assume the samples were gathered independently. Let $\overline{X} = \frac{1}{n}\sum_{i=1}^{n_1} X_i$ and $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n_2} Y_i$ be the sample means. Let $\hat{\Delta} = \overline{X} - \overline{Y}$. Since $\overline{X}$ is an unbiased estimator for μ_1 and $\overline{Y}$ is an unbaised estimator of μ_2 (by Theorem 2.8.A) then $\hat{\Delta}$ is an unbiased estimator of Δ. We take $\hat{\Delta} - \Delta$ as the "pivot random variable." Since the samples are independent,

$$\begin{aligned} \text{Var}(\hat{\Delta}) &= \text{Var}(\overline{X} - \overline{Y} = E((\overline{X} - \overline{Y})^2) - (\mu_1 - \mu_2)^2 \text{ by Note 1.9.A} \\ &= E(\overline{X}^2 - 2\overline{XY} + \overline{Y}^2) - \mu_1^2 + \mu_1\mu_2 - \mu_2^2 \end{aligned}$$

$$\begin{aligned}
&= \left(E(\overline{x}^2) - \mu_1^2\right) + \left(E(\overline{Y}^2) - \mu_2^2\right) - 2E(\overline{XY}) + 2\mu_1\mu_2 \\
&= \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) - 2(E(\overline{XY}) - \mu_1\mu_2) \text{ by Note 1.9.A} \\
&= \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) - 2E(\overline{X})E(\overline{Y}) - \mu_1\mu_2) \text{ by Theorem 2.4.4,} \\
&\quad \text{since } \overline{X} \text{ and } \overline{Y} \text{ are independent} \\
&= \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) - 2(\mu_1\mu_2 - \mu_2\mu_2) \\
&= \text{Var}(\overline{X}) + \text{Var}(\overline{Y}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2} \text{ by Theorem 2.8.A.}
\end{aligned}$$

Let $S_1 = \dfrac{1}{n_1 - 1}\displaystyle\sum_{i=1}^{n_1}(X_i - \overline{X})^2$ and $S_2 = \dfrac{1}{n_2 - 1}\displaystyle\sum_{i=1}^{n_2}(Y_i - \overline{Y})^2$ be the sample variances. Estimating the population variances with these sample variances, gives the random variable $Z = \dfrac{\hat{\Delta} - \Delta}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$. By the Central Limit Theorem (Theorem 4.2.1), A has an approximate standard normal distribution. So the approximate $(1 - \alpha)100\%$ confidence interval for $\Delta = \mu_1 - \mu_2$ is

$$\left((\overline{x} - \overline{y}) - z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, (\overline{x} - \overline{y}) + z_{\alpha/2}\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}\right).$$

This is the *large sample* $(1 - \alpha)100\%$ *confidence interval* for $\mu_1 - \mu_2$. Quantity $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ is the *standard error* of $\overline{X} = \overline{Y}$.

Note. Similar to the previous note, we now consider random variables A and Y which are both normally distributed and with the same variance (so $\sigma_1^2 = \sigma_2^2 = \sigma^2$). So the distributions of X and Y are the same shape; they only differ in mean. This is called a *location model*. So we consider X with a $N(\mu_1, \sigma)$ distribution, Y with a $N(\mu_2, \sigma)$ distribution, $X_1, X_2, \ldots, X_{n_1}$ a random sample on X, and $Y_1, Y_2, \ldots, Y_{n_2}$ a random sample on Y, where the samples are independent. Let $n = n_1 + n_2$. Again, we take $\hat{\Delta} = \overline{X} - \overline{Y}$. Norm $\overline{X}$ is distributed $N(\mu_1, \sigma^2/n_1)$ and $\overline{Y}$ is distribution $N(\mu_2, \sigma^2/n_2)$ by Theorem 2.8.A. By Theorem 3.4.2, $\overline{X} - \overline{Y}$ has mean $\mu_1 - \mu_2$ and variance $\sigma^2/n_1 + \sigma^2/n_2$. Therefore $\dfrac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ has a $N(0,1)$ distribution. Let

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

Since $E[S_1^2] = E[S_2^2] = \sigma^2$ by Theorem 2.8.A, then $E[S_p^2] = \sigma^2$, so S_p^2 is an unbiased estimation of σ^2. S_p^2 is called the pooled estimator of σ^2. By Theorem 3.6.1(c) (Student's Theorem), $(n_1 - 1)S_1^2/\sigma^2$ had a $\chi^2(n_1 - 1)$ distribution and $(n_2 - 1)S_2^2/\sigma^2$ has a $\chi^2(n_2 - 1)$ distribution. Since S_1^2 and S_2^2 are independent (since X and Y are) then by Corollary 3.3.1, $(n-2)S_p^2/\sigma^2$ has a $\chi^2(n-2)$ distribution. By Exercise 3.5.B, $\overline{X}$ and S_1^2 and $\overline{X}$ are independent. In Note 3.6.A, we showed that for independent random variables X and V where W is $N(0,1)$ and V is $\chi^2(r)$, the random variable $T = W/\sqrt{V/r}$ has a t-distribution with r degrees of freedom. Taking

$$W = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \text{ and } V = (n-2)\frac{S_p^2}{\sigma^2}$$

where $r = 2$, we have

$$T = \frac{((\overline{x} - \overline{Y}) - (\mu_1 - \mu_2))/(\sigma\sqrt{n_1^{-1} + n_2^{-1}})}{\sqrt{((n-2)S_p^2/\sigma^2)/(n-2)}} = \frac{(\overline{X} - \overline{Y}) - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

has a t-distribution with $r = 2$ degrees of freedom. So the $(1-\alpha)100\%$ confidence interval for $\Delta = \mu_1 - \mu_2$ in this case is

$$\left((\overline{x} - \overline{y}) - t_{\alpha/2,n-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}, (\overline{x} - \overline{y}) + t_{\alpha/2,n-2}s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\right).$$

In Exercise 4.2.23, the problem of finding a confidence interval for the difference

$\mu_1 - \mu_2$ between two means of two normal distributions if the variances σ_1^2 and σ_2^2 are known but not necessarily equal, is to be discussed.

Note. Let X and Y be two independent random variables with Bernoulli distributions where the probability of success for X is p_1 and the probability of success for Y is p_2 (that is, X and Y have the binomial distribution $b(1, p_1)$ and $b(1, p_2)$, respectively). Let $X_1, X_2, \ldots, X_{n_1}$ be a random sample on X and let $Y_1, Y_2, \ldots, Y_{n_2}$ be a random sample on Y. Assume the samples are independent of one another and let $n = n_1 + n_2$ be the total sample size. We approximate the propositions with the sample proportions $\hat{p}_1$ and $\hat{p}_2$, and approximate the variances with the sample variances $\hat{\sigma}_1^2 = \hat{p}_1(1 - \hat{p}_1)$ and $\hat{\sigma}_2^2 = \hat{p}_2(1 - \hat{p}_2)$. Then an approximate $(1 - \alpha)100\%$ confidence interval for $p_1 - p_2$ is

$$\left((\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}, (\hat{p}_1 - \hat{p}_2) - z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}} \right),$$

similar to above.

Example 4.2.5. In the original clinical study of the Jonas Salk polio vaccine in 1954, a group of children were randomly assigned to two categories. A group of 200,745 children received the vaccine ("Treated") and a group of 201,229 children did not receive the vaccine but instead got a placebo ("Control"). Let p_C and p_T denote the proportions of polio cases for the control and treated groups, respectively. The data is:

Group	**# Children**	**# Polio Cases**	**Sample Proportion**
Treated	200,745	57	0.000284
Control	201,229	199	0.00989

(Notice that Hogg, McKean, and Craig have a typographical error in the Control sample proportion.) So $0.000989 = \hat{p}_C > \hat{p}_T = 0.00284$. For a 95% confidence interval, we take $\alpha = 0.05$ and $z_{\alpha/2} = z_{0.025} = 1.96$. This gives the interval for $p_C - p_T$:

$$\left((0.000989 - 0.000284) - 1.96\sqrt{\frac{0.000989(1-0.000989)}{201{,}229} + \frac{0.000284(1-0.00284)}{200{,}745}}, \right.$$

$$\left. (0.000989 - 0.000284) + 1.96\sqrt{\frac{0.000989(1-0.000989)}{201{,}229} + \frac{0.000284(1-0.00284)}{200{,}745}} \right)$$

$$= (0.000549, 0.000861).$$

Here, we have carried computations to three significant digits. Notice that Hogg, McKean, and Craig give the interval $(0.00054, 0.00087)$, a larger interval than the one computed here. The reason for this difference is unclear…it may simply be roundoff error but it also could have something to do with the software used by the textbook authors (or there may be other typographical errors in the data presented). Since $0 \not\in (0.000549, 0.000861)$, then we are 95% confident that $p_C > p_T$ and that the vaccine reduces the incidence of polio. This foreshadows how we will use confidence intervals in hypothesis testing Section 4.5. Introduction to Hypothesis Testing.

Revised: 8/4/2021