STAT 515 Lec 10

Confidence intervals for the mean and proportion

Karl Gregory

Confidence intervals

Now we come to the payoff. The goal of statistics is to learn about a population from a random sample. We here concern ourselves with the questions:

1. What can we say about μ based on $\bar{X}$?
2. What can we say about p based on $\hat{p}$?

We use $\bar{X}$ to estimate μ, but we know that if we took another random sample, we would not get the same value of $\bar{X}$. Likewise, we use $\hat{p}$ to estimate p, but we know that if we took another random sample, it is unlikely that we would get the same $\hat{p}$. It would be silly to say, "We believe that the value of μ is equal to $\bar{X}$," when $\bar{X}$ is the mean of a random sample. Likewise, it would be silly to say, "We believe that the value of p is equal to $\hat{p}$," when $\hat{p}$ is the proportion of successes in a random sample. What then can we say? Instead of saying that μ is equal to $\bar{X}$ or that $\hat{p}$ is equal to p, we say, "We are fairly confident that μ lies within some interval around $\bar{X}$," or, "We are fairly confident that p lies within some interval around $\hat{p}$." Such an interval is called a *confidence interval*: it is an interval constructed from the random sample such that it will contain the parameter of interest, be it μ or p, with a certain probability.

CI for the mean of a Normal population (σ known)

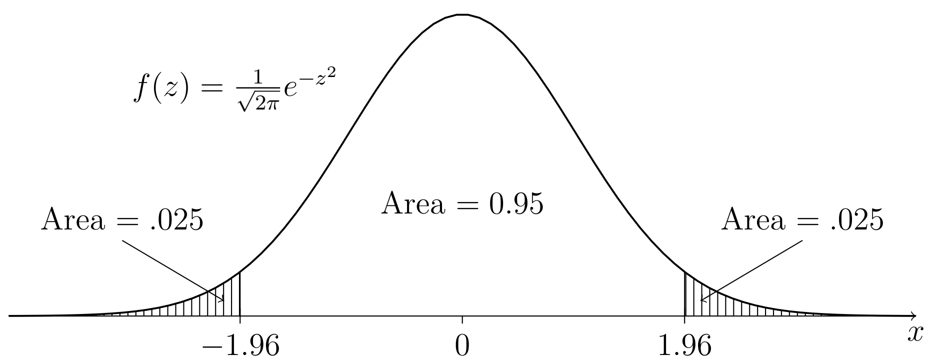
Suppose we draw a random sample $X_1, \dots, X_n$ from a population with an unknown mean μ and a known variance σ^2. Suppose in addition that $\bar{X}$ behaves like a Normal$(\mu, \sigma^2/n)$ random variable (i.e. the population is Normal or the sample size n is large enough). We would like to construct an interval (L, U), where L and U are computed from the sample, such that $P(L \leq \mu \leq U)$ is large. That is, we would like our interval (U, L) to contain the value of μ with high probability.

To start off, let's say we want an interval which contains μ with probability 0.95. We may arrive at such an interval in two steps:

1. First note that if $\bar{X}$ has the Normal$(\mu, \sigma^2/n)$ distribution, then

$$P\left(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96\right) = 0.95.$$

   Why? Because $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is a Normal$(0, 1)$ random variable, and the area under the Normal$(0, 1)$ density function between -1.96 and 1.96 is 0.95:



2. We may rearrange the above expression to get

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = .95.$$

   So the desired interval is given by

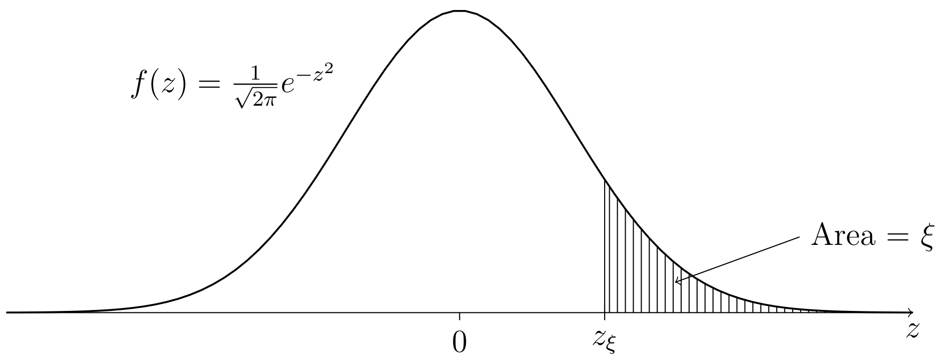
$$\bar{X} \pm 1.96\frac{\sigma}{\sqrt{n}}.$$

We call this a 95% *confidence interval for* μ.

More generally, for any $\alpha \in (0, 1)$, we consider the construction of $(1 - \alpha)100\%$ confidence intervals, where the value α is the probability that our confidence interval will *not* contain μ. For a 95% confidence interval, the corresponding value of α is 0.05. We refer to $1 - \alpha$ as the *confidence level* of the confidence interval. To give a general expression for a $(1-\alpha)100\%$ confidence interval for μ, we define for any $0 < \xi < 1$ the quantity z_ξ as the value such that

$$P(Z > z_\xi) = \xi,$$

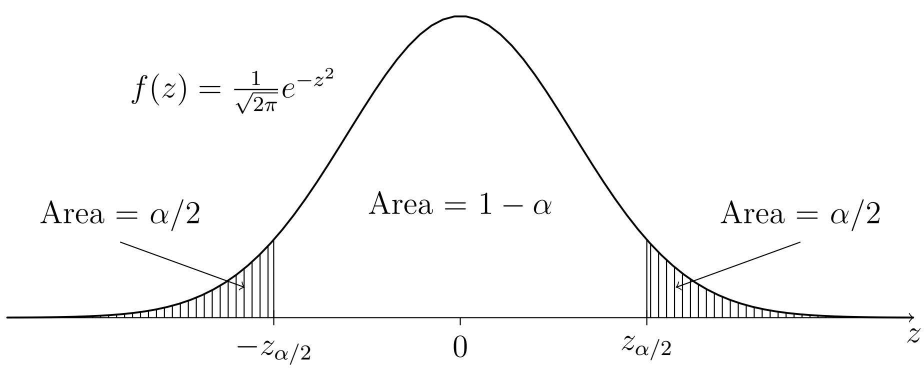
where Z is a random variable having the Normal$(0, 1)$ distribution. The value z_ξ admits the depiction below:



Now, if $\bar{X}$ has the Normal$(\mu, \sigma^2/n)$ distribution, we may construct for any $\alpha \in (0,1)$ a $(1-\alpha)100\%$ confidence interval for the mean μ by noting that

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha,$$

corresponding to the picture



We may rearrange the above expression to get

$$P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha,$$

from which we can see that a $(1-\alpha)100\%$ confidence interval for μ may be constructed as

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}.$$

We now state this formally:

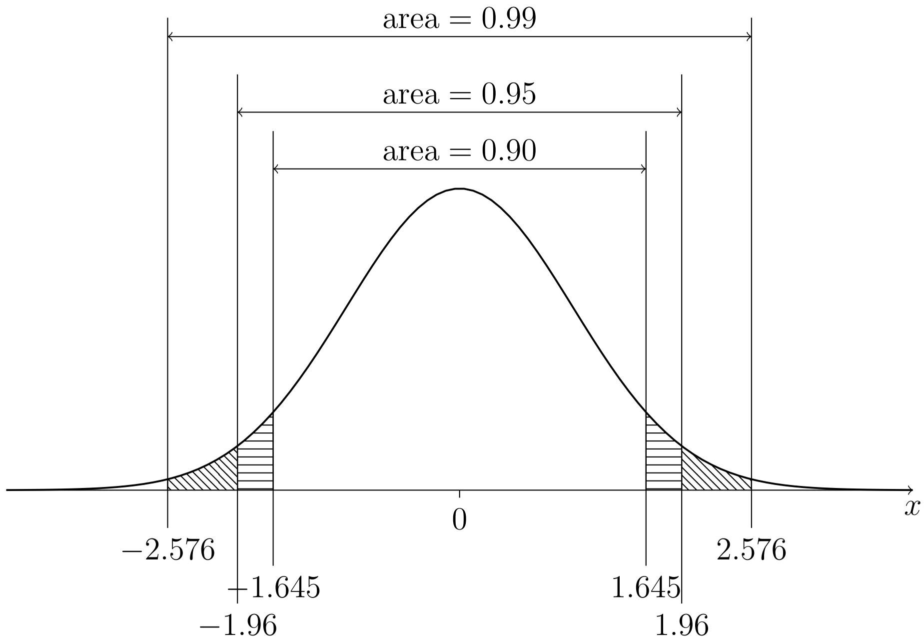
Result: Confidence interval for mean of Normal population with σ known

For a random sample $X_1, \ldots, X_n \overset{\text{ind}}{\sim}$ Normal(μ, σ^2) with σ known,

$$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is a $(1-\alpha) \times 100\%$ confidence interval for μ.

We are very often interested in building confidence intervals at the 0.90, 0.95, or 0.99 confidence levels, for which the following diagram depicts the necessary quantiles, $z_{0.05} = 1.645$, $z_{0.025} = 1.96$, and $z_{0.005} = 2.576$, respectively, of the standard Normal distribution:



Exercise. Suppose you have a random sample of size 35 with sample mean $\bar{X} = 25$ from a right-skewed population with unknown μ and with $\sigma^2 = 10$. What is a 90% confidence interval for the mean, and what is its interpretation?

Answer: The population distribution is not Normal, but since the sample size is greater than 30, we can treat $\bar{X}$ like a Normal$(\mu, 10/35)$ random variable. For a 90% confidence interval we have $\alpha = 0.10$, so we need $z_{\alpha/2} = z_{0.05}$. We have $z_{0.05} = 1.645$. Therefore, a 90% confidence interval for μ is given by

$$25 \pm 1.645\frac{\sqrt{10}}{\sqrt{35}} = (24.12, 25.88).$$

We are 90% confident that the mean μ lies within the interval $(24.12, 25.88)$.

Exercise. Suppose you have a random sample of size 8 with sample mean $\bar{X} = 12$ from a population with a Normal distribution with unknown μ and with $\sigma^2 = 9$. What is a 95% confidence interval for the mean and what is its interpretation?

Answer: Since the population is Normal, $\bar{X}$ has the Normal$(\mu, 9/8)$ distribution even though the sample size is small. For a 95% confidence interval we have $\alpha = 0.05$, so we need to get $z_{\alpha/2} = z_{0.025}$. We find from the Z table that $z_{0.025} = 1.96$. Therefore, a 95%

confidence interval for μ is given by

$$12 \pm 1.96 \frac{3}{\sqrt{8}} = (9.92, 14.08).$$

We are 95% confident that the mean μ lies within the interval $(9.92, 14.08)$.

Note: The textbook calls $1-\alpha$ the *confidence coefficient*.

CI for the mean of a non-Normal population (σ known)

when the population is not Normally distributed, the sample mean $\bar{X}_n$ does not have a Normal distribution, so the confidence interval for the mean given the previous section cannot be used; however, according to the central limit theorem, the behavior of the quantity

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

becomes more and more like that of a standard Normal random variable as the sample size n gets larger. So if n is large, then the confidence interval given in the previous section will still be approximately correct. We state this here as a result:

Result: Confidence interval for mean of a non-Normal pop. with σ known

Let $X_1, \ldots, X_n$ be a random sample from a non-Normal distribution with mean μ and variance $\sigma^2 < \infty$. Then

$$\bar{X}_n \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

contains μ with probability closer and closer to $1-\alpha$ for larger and larger n.

Confidence interval for the proportion p

We construct a $(1-\alpha)100\%$ confidence interval for p based on $\hat{p}$ in much the same way as we did for μ based on $\bar{X}$. Recall that $\hat{p}$ is nothing but the mean of a random sample of size n of the Bernoulli(p) random variables

$$X_i = \begin{cases} 1 & \text{if outcome } i \text{ a "success"} \\ 0 & \text{if outcome } i \text{ a "failure"} \end{cases} \qquad \text{for } i = 1, \ldots, n,$$

where the probability of "success" is p and the probability of "failure" is $1-p$. If the conditions $np \geq 5$ and $n(1-p) \geq 5$ are satisfied, then the central limit theorem tells us that

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \text{ approximately follows the Normal}(0,1) \text{ distribution,}$$

which allows us to write

$$P\left(-z_\alpha < \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} < z_\alpha\right) \approx 1-\alpha.$$

Rearranging the above gives

$$P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right) \approx 1-\alpha,$$

from which we see that an approximate $(1-\alpha)100\%$ confidence interval for p could be constructed as

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}. \tag{1}$$

However, we cannot compute this interval because we don't know the value of p. There are a couple of ways to deal with this.

The Wald interval (for $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$)

Our first instinct may be to replace p in (1) by its estimator $\hat{p}$. We certainly may do this, but the resulting interval is not very reliable unless the sample size is very large; it is especially unreliable when the true proportion is close to 0 or 1.

When the sample size is very large and it is believed that p is not very close to 0 or 1 (we can check the condition $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$), an approximate $(1-\alpha)100\%$ confidence interval for p can be constructed by

$$\hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

Result: Wald interval for a population proportion

For a random sample $X_1, \ldots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$,

$$\hat{p}_n \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

is an approximate $(1-\alpha) \times 100\%$ confidence interval for p.

Use only when $n\hat{p} \geq 15$ and $n(1-\hat{p}) \geq 15$.

This is called the *Wald interval*, and its performance is notoriously bad unless n is very large. By "bad peformance", we mean that the actual probability that the interval contains the true value of p is quite different from the specified probability of $1-\alpha$.

The Agresti-Coull interval (for $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$)

It has been shown that the following interval works much better than the Wald interval: Define

$$\tilde{p} = \frac{\#\{\text{successes}\} + 2}{n+4}.$$

Now, a much better $(1-\alpha)100\%$ confidence interval for p can be constructed by

$$\tilde{p} \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}.$$

This is called the *Agresti-Coull* interval, and it has been shown to have good performance provided $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$, so it can be used under much smaller sample sizes than the Wald interval. The textbook calls this interval the Wilson adjusted interval.

Result: Agresti-Coull interval for a population proportion

For a random sample $X_1, \dots, X_n \overset{\text{ind}}{\sim} \text{Bernoulli}(p)$,

$$\tilde{p}_n \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n+4}}, \qquad \text{where } \tilde{p} = \frac{\#\{\text{successes}\} + 2}{n+4},$$

is an approximate $(1-\alpha) \times 100\%$ confidence interval for p.

Use only when $n\hat{p} \geq 5$ and $n(1-\hat{p}) \geq 5$.

Exercise. Suppose you draw a random sample of 1000 registered voters. Suppose that 478 of the 1000 say they will vote for candidate A. Build a 95% confidence interval for the proportion of registered voters who will vote for candidate A.

Answer: For a 95% confidence interval, $\alpha = 0.05$, and $z_{0.025} = 1.96$. So the Agresti-Coull interval is given by

$$\frac{480}{1004} \pm 1.96\sqrt{\frac{(480/1004)(1-480/1004)}{1004}} = 0.478 \pm 0.031 = (0.447, 0.509)$$

The Wald interval is in this case the same out to three decimal places because the sample size is so large:

$$\frac{478}{1000} \pm 1.96\sqrt{\frac{(478/1000)(1-478/1000)}{1000}} = 0.478 \pm 0.031 = (0.447, 0.509)$$

Exercise. Suppose you randomly sample 50 USC undergraduates and find that 5 of them hang-dry their laundry to conserve electricity.

1. Build a 95% Agresti-Coull confidence interval for p, the proportion of USC undergraduates who hang-dry their laundry to save electricity.

   Answer: For a 95% confidence interval $\alpha = 0.05$, so we use $z_{0.025} = 1.96$. Then the Agresti-Coull interval is

$$\frac{7}{54} \pm 1.96\sqrt{\frac{(7/54)(1-7/54)}{54}} = (0.040, 0.219)$$

2. Build a 95% Wald confidence interval for p.

   Answer: The Wald interval is

$$\frac{5}{50} \pm 1.96\sqrt{\frac{(5/50)(1-5/50)}{50}} = (0.017, 0.183)$$

   The Wald and Agresti-Coull intervals in this example are very different. It is better here to use the Agresti-Coull interval because of the smaller sample and the small number of successes.