Confidence Interval Construction: Examples

Peter E. Freeman - Statistics & Data Science - Carnegie Mellon University

June 2024

The Setup

Let's assume that we sample n independent and identically distributed (iid) data from a distribution with parameter θ, and let Y be a statistic formed from these data (e.g., $\bar{X}$). (Y is typically a sufficient statistic with a known distribution, but it need not be.) To determine an interval bound, we solve the following equation for θ:

$$F_Y(y_{\rm obs}|\theta) - q = 0\,,$$

where

- $F_Y(\cdot)$ is the cumulative distribution function for Y
- $y_{\rm obs}$ is the observed statistic value
- q is an appropriate quantile

To determine the value for the appropriate quantile, we need to know two things: the type of interval we trying to construct (two-sided? one-sided lower bound? one-sided upper bound?), and whether the expected value of the adopted statistic, $E[Y]$, increases with θ (i.e., increases in value as θ increases in value), or decreases. Given those two pieces of information, we can pull the appropriate quantile value off of the following reference table

Interval Type	$E[Y]$ Increases With θ?	q for Lower Bound	q for Upper Bound
two-sided	yes	$1-\alpha/2$	$\alpha/2$
	no	$\alpha/2$	$1-\alpha/2$
one-sided lower	yes	$1-\alpha$	$-$
	no	α	$-$
one-sided upper	yes	$-$	α
	no	$-$	$1-\alpha$

Example 1: An Analytic Solution

We sample a single datum with value $X = x_{\rm obs} = 1$ from an exponential distribution with probability density function

$$f_X(x) = \frac{1}{\theta} e^{-x/\theta}\,.$$

What is a 95% lower-bound for θ?

1. Identify an appropriate statistic. That's trivial here: $Y = X$ (and hence $y_{\rm obs} = x_{\rm obs} = 1$).
2. Determine the cdf for the random variable Y. Stated without proof, that's

$$F_Y(y|\theta) = 1 - e^{-y/\theta}\,.$$

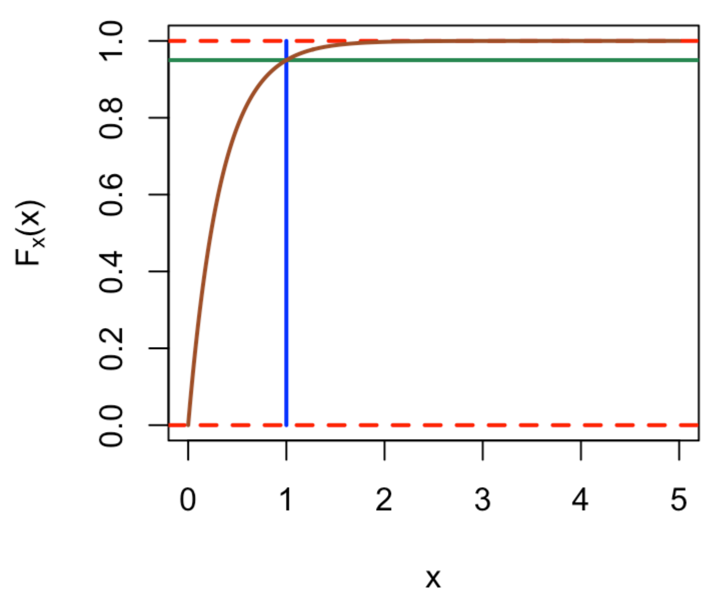
3. The expected value of Y is $E[Y] = \theta$. As θ increases, $E[Y]$ increases.
4. We want a one-sided lower bound (with $\alpha = 0.05$) and, according to (3), we are on the "yes" line. Thus $q = 1 - \alpha = 0.95$.

We now have all the pieces necessary to derive the lower bound:

$$
\begin{aligned}
F_Y(y_{\rm obs} \vert \hat{\theta}_L) - q &= 0 \\
\Rightarrow ~~~ 1 - e^{-y_{\rm obs}/\hat{\theta}_L} - 0.95 &= 0 \\
\Rightarrow ~~~ e^{-y_{\rm obs}/\hat{\theta}_L} &= 0.05 \\
\Rightarrow ~~~ -y_{\rm obs}/\hat{\theta}_L &= \log(0.05) \\
\Rightarrow ~~~ \hat{\theta}_L &= -\frac{y_{\rm obs}}{\log(0.05)} = 0.334 \,.
\end{aligned}
$$

Below, we show how the cdf for the random variable Y, assuming $\hat{\theta}_L = 0.334$, passes through the intersection of the blue line (representing $y_{\rm obs}$) and the green line (representing $q = 0.95$). Plots like this are good to show to students: if the observed value changes, the blue line shifts, and thus we have to change $\hat{\theta}_L$ to get the cdf to once again pass through the intersection of the lines. In short: confidence intervals are random intervals.

```
y.obs <- 1
plot(c(y.obs,y.obs),c(0,1),typ="l",xlab="x",ylab=expression(F[x]*"(x)"),col="blue",xlim=
c(0,5),lwd=2)
abline(h=0,col="red",lwd=2,lty=2)
abline(h=1,col="red",lwd=2,lty=2)
abline(h=0.95,col="seagreen",lwd=2)
theta   <- 0.334
x.plot  <- seq(0,5,by=0.01)
Fx.plot <- 1 - exp(-x.plot/theta)
lines(x.plot,Fx.plot,col="sienna",lwd=2)
```



Example 2: A Quasi-Analytic Solution

We sample $n = 10$ iid data from a normal distribution with mean μ and known variance $\sigma^2 = 4$. We observe $\bar{X} = 5$. What is a 95% upper bound on μ?

1. Is $\bar{X}$ an appropriate statistic? Yes: it is a sufficient statistic (as found via likelihood factorization) and we know the sampling distribution for $\bar{X}$ (as found via the method of moment-generating functions):

$$\bar{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) .$$

2. The cdf for the random variable $Y = \bar{X}$ is

$$F_Y(y|\mu, \sigma^2) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{y-\mu}{\sqrt{2}\sigma}\right)\right] .$$

3. We know that $E[Y] = E[\bar{X}] = \mu$, so $E[Y]$ does increase with μ.
4. We want a one-sided upper bound (with $\alpha = 0.05$) and, according to (3), we are on the "yes" line. Thus $q = \alpha = 0.05$.

So let's solve!

$$\begin{aligned}
F_Y(y_{\rm obs}|\hat{\mu}_U, \sigma^2/n) - q &= 0 \\
\Rightarrow \quad \frac{1}{2}\left[1 + \text{erf}\left(\frac{\sqrt{n}(y_{\rm obs} - \hat{\mu}_U)}{\sqrt{2}\sigma}\right)\right] - 0.05 &= 0 \\
\Rightarrow \quad 1 + \text{erf}\left(\sqrt{n}(\frac{y_{\rm obs} - \hat{\mu}_U)}{\sqrt{2}\sigma}\right) &= 0.1 \\
\Rightarrow \quad \text{erf}\left(\frac{\sqrt{n}(y_{\rm obs} - \hat{\mu}_U)}{\sqrt{2}\sigma}\right) &= -0.9 \\
\Rightarrow \quad \frac{\sqrt{n}(y_{\rm obs} - \hat{\mu}_U)}{\sqrt{2}\sigma} &= \text{erf}^{-1}(-0.9) \\
\Rightarrow \quad y_{\rm obs} - \hat{\mu}_U &= \sqrt{2}\frac{\sigma}{\sqrt{n}}\text{erf}^{-1}(-0.9) \\
\Rightarrow \quad \hat{\mu}_U &= y_{\rm obs} - \sqrt{2}\frac{\sigma}{\sqrt{n}}\text{erf}^{-1}(-0.9) = 6.040 .
\end{aligned}$$

To compute the inverse error function, we use the `erfinv()` function of `R`'s `pracma` package.

Does this match what we would derive using a canned formula from introductory statistics?

$$\bar{X} + z_{0.95}\frac{\sigma}{\sqrt{n}} = 5 + 1.645\frac{2}{\sqrt{10}} = 6.040 ,$$

where $z_{0.95}$ = `qnorm(0.95)` = 1.645. Yep…our result matches perfectly.

Example 3: A Numerical Solution

Let's repeat Example 2, but using `R`'s `uniroot()` function.

First, we need to define a function that returns the value of $F_Y(y|\theta) - q$:

```
f <- function(mu,y.obs,q,sigma,n)
{
  pnorm(y.obs,mean=mu,sd=sigma/sqrt(n)) - q
}
```

Note that what we want to solve for (`mu`) has to be the first argument to `f` .

Second, we pass our new function to `uniroot()` along with a range of values over which to search for the one root of the equation:

```
uniroot(f,interval=c(-1000,1000),y.obs=5,q=0.05,sigma=2,n=10)$root
```

```
## [1] 6.040307
```

Done. Note that since μ can take on any value, we define a large range of negative and positive numbers over which to search. Also note that if σ^2 is unknown, we can utilize the t distribution and write, e.g.,

```
f <- function(mu,y.obs,q,S,n)
{
  pt((y.obs-mu)/(S/sqrt(n)),n-1) - q
}
uniroot(f,interval(-1000,1000),y.obs=5,q=0.05,S=[sample sd],n=10)$root
```

where S is the sample standard deviation.

Example 4: But Does This Work with Discrete Distributions?

The short answer: yes. The key is that the parameter θ is, in typical situations, continuously valued, and so the basic algorithm doesn't change.

Let's conduct an experiment in which we flip a coin $k = 10$ times. We record the number of heads. We then repeat the experiment such that we have $n = 20$ outcomes, and we find that $\bar{X} = 116/20 = 5.8$. What is a two-sided confidence interval for the success probability p?

1. It turns out that $\bar{X}$ is not an appropriate statistic here, because we cannot easily write down its sampling distribution. (We can define it exactly numerically, but…) On the other hand,

$$n\bar{X} = \sum_i X_i \sim \text{Binomial}(nk, p)\,,$$

   as one can easily derive using the method of moment-generating functions. Hence we'll use $Y = \sum_i X_i$, with the observed value $y_{\rm obs} = 116$.
2. The cdf is given in (1).
3. We know that $E[Y] = nkp$ increases with p.
4. Thus the lower bound is associated with the value $q = 0.975$ and the upper bound is associated with the value $q = 0.025$.

Let's solve!

```
f <- function(p,y.obs,q,nk)
{
  pbinom(y.obs,size=nk,prob=p) - q
}
uniroot(f,interval=c(0,1),y.obs=116,nk=200,q=0.975)$root # lower bound
```

```
## [1] 0.5133761
```

```
uniroot(f,interval=c(0,1),y.obs=116,nk=200,q=0.025)$root # upper bound
```

```
## [1] 0.6492334
```

Hmm…it appears the coin may very well be an unfair one, as $p = 0.5$ falls outside the interval.

Example 5: What to Do When All Hope is Lost

AKA, Working with Beta Distributions

We sample $n = 5$ iid data from a Beta$(a, 2.6)$ distribution, with the observed data being $\{0.3313, 0.1908, 0.1089, 0.0006937, 0.1642\}$. What is a 95% one-sided upper bound for a?

When we apply likelihood factorization, the sufficient statistic $U = \prod_{i=1}^{n} X_i$ pops out…but we don't necessarily want to use this in general computations, because products can blow up quickly. (Also, we don't know the sampling distribution for U.) Recalling that functions of sufficient statistics are themselves sufficient statistics, we define $Y = -\log U = -\sum_{i=1}^{n} \log X_i$. We still don't know the sampling distribution for this statistic, but it *is* a summation and thus numerically easier to work with. (And why the minus sign? So that the value of Y is a positive number…that's really the only reason.)

For a beta distribution, $E[X] = a/(a+b)$ increases as a increases (with b fixed), so…

- $E[\log(X)]$ also increases as a increases, so…
- $E[\sum_i \log(X_i)]$ also increases as a increases, so…
- $E[-\sum_i \log(X_i)]$ *decreases* as a increases.

This means that we are on the upper bound/"no" line of the reference table, and thus that $q = 1 - \alpha = 0.95$.

What now? We don't know the sampling distribution for Y. The jig is up. Game over. [Insert your own cliche here.]

What we can do is simulate the empirical cdf for Y, given a. We do this in the code chunk below. First, we define a function `f` inside of which we

1. create `num.sim` separate datasets, each of length `n`, given parameter values for `a` and `b`;
2. compute the statistic Y for each dataset;
3. determine the proportion of Y values that are less than `y.obs` (this is the empirical cdf $F_Y(y_{\rm obs} \vert a, b)$); and
4. return the value $[F_Y(y_{\rm obs} \vert a, b) - q]^2$.

Instead of passing `f` to `uniroot()`, we pass it to `optimize()`, with the idea that we are trying to find the value of a that minimizes the objective function $(F_Y(y_{\rm obs} \vert a, b) - q)^2$. (Hence the squaring of the quantity!)

```r
f <- function(a,n,b,y.obs,q,num.sim=100000,seed=236)
{
  set.seed(seed)
  X <- matrix(rbeta(n*num.sim,shape1=a,
                    shape2=b),nrow=num.sim)
  Y <- apply(X,1,function(x){-sum(log(x))})
  (sum(Y <= y.obs)/num.sim-q)^2
}
# c(0.01,2) is a user-set range over which to explore
print(optimize(f,c(0.01,2),n=5,b=2.6,y.obs=14.058,q=0.95)$minimum)
```

```
## [1] 0.9292644
```

We find that the approximate 95% upper bound on a is roughly 0.93.

So: even when we *don't* know the sampling distribution of our adopted statistic, we can *still* estimate interval bounds! Huzzah!