Precision of a Sampling Distribution’s Summary Statistics

Teddy Weaver

Contents

Random sampling is used to **estimate** summary statistics of a larger population as measuring the entire population is often impracticle or downright impossible. This process creates a probability distribution of the sampling statistics, also called the sampling distribution.

A side-effect of random sampling is that it causes uncertainty – this is why it results in a distribution and not a single value. This blog post dives into this uncertainty to show that certain quantiles can be estimated more accurately than others and how this depends to the distribution.

The Setup

Here are the included parameters and brief description. Expand.grid is being used to create a matrix of all combinations. Each row can then be passed as a set of parameters, allowing us to iterate through each combination of settings.

```r
p_seq <- seq(.05,.95, .05) # Probability intervals

sim_settings <- expand.grid(
  N = 200, # Sample Size
  M = 5000, # Number of Samples
  D = c("norm", "exp", "f3", "f4"), # Distributions
  KEEP.OUT.ATTRS = FALSE,
  stringsAsFactors = FALSE
)
```

The settings also include an array of the four distributions used as examples. Both f3 and f4 are mixed distributions that can described by the their respective density functions below.

```r
df3 <- function(x){
  .5*dnorm(x) + .3*dnorm(x,4) + .2*dnorm(x,-4,2)
}

df4 <- function(x) {
  .5*dbeta(x,5,1) + .5*dbeta(x,1,5)
}
```

The How

To visualize uncertainty, we'll be finding the sampling distribution of the length of the middle 95% – or the difference between the 2.5% and 97.5% percentiles. We can think of this summary statistic as the width of the distribution, excluding outliers in a hypothetical data set.

We'll also be taking a porgrammatic approach to solving this problem. As a result, some helper functions have been created to return the various required distribution functions.

```
get_dist <- function(D) {
  return(
    list("rand" = get("r" %|% D),
         "quant" = get("q" %|% D),
         "dens" = get("d" %|% D),
         "dist" = get("p" %|% D))
    )
}
```

To find the mid-95% length for each of the 19 probability values in `p_seq` we used random distribution functions to create 5,000 samples of size N (200). The length was then found by subtracting the .025 quantile from the .975 quantile.

In addition, quantile and density values were found for each probability value using their respective distribution functions. This allows us to easily compare the mid-95% length to the density and cumulative distribution functions.

```
calc_length <- function(D, N, M, p) {
  len_p <- length(p)

  sim_quantile <- array(NA, dim = c(M,len_p))
  sim_length <- vector(length=len_p)

  set.seed(1)
  for(i in 1:M){
    sim_quantile[i,] <- D[[1]](N) %>% quantile(probs = p)
  }

  for(i in 1:len_p) {
    quant <- quantile(sim_quantile[,i], c(.025, 0.975))
    sim_length[i] <- diff(quant)
  }
  return(sim_length)
}

# Derive Length and Density of the function. Output into dataframe.
main <- function(params, p_seq) {
  inputs <- as.list(params)

  with(inputs, {
    dist.type <- get_dist(D)
    s.length <- calc_length(dist.type["rand"], N, M, p_seq)
    s.quantile <- p_seq %>% (dist.type["quant"][[1]])
    s.density <- s.quantile %>% (dist.type["dens"][[1]])
    return(
      data.frame("prob" = p_seq,
                 "dist" = D,
                 "N" = N,
                 "M" = M,
                 "length" = s.length,
                 "density" = s.density,
                 "quantile" = s.quantile
                 )
      )
  })
}
```

This process is done for each of our settings combinations in the `sim_settings` data frame. The results data frame from each iteration is added to a list then combined into a final data frame using `rbindlist`. Below is a sample of the combined results dataframe.

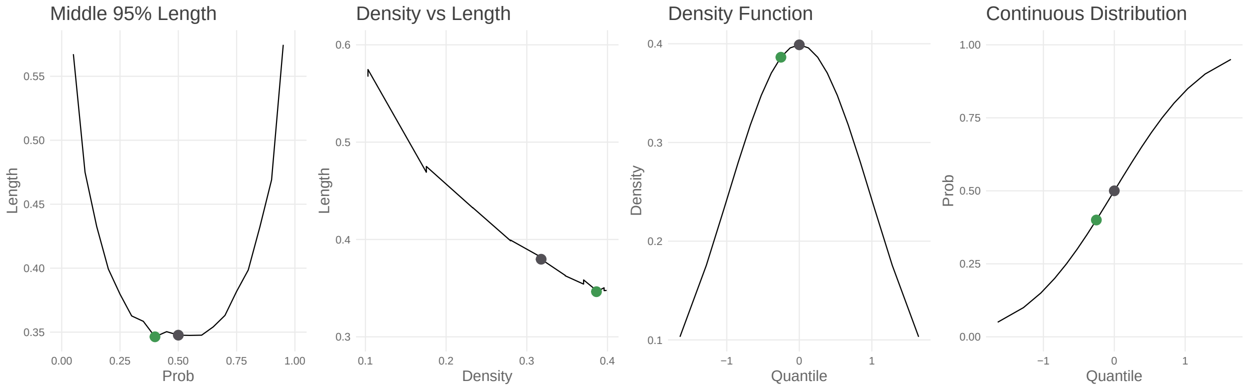
prob	dist	N	M	length	density	quantile
0.05	norm	200	5000	0.5672967	0.1031356	-1.6448536
0.10	norm	200	5000	0.4750435	0.1754983	-1.2815516
0.15	norm	200	5000	0.4326273	0.2331588	-1.0364334
0.20	norm	200	5000	0.3993222	0.2799619	-0.8416212
0.25	norm	200	5000	0.3796905	0.3177766	-0.6744898

The Results

As one might expect the shortest length, or location of most greatest accuracy, occurs during the densest region of the distribution - also the steepest slope on the CDF. Below are various graphs for each distribution to illustrate how the quantiles with more precision (shortest length) compare to the median, with both called out on each graph.

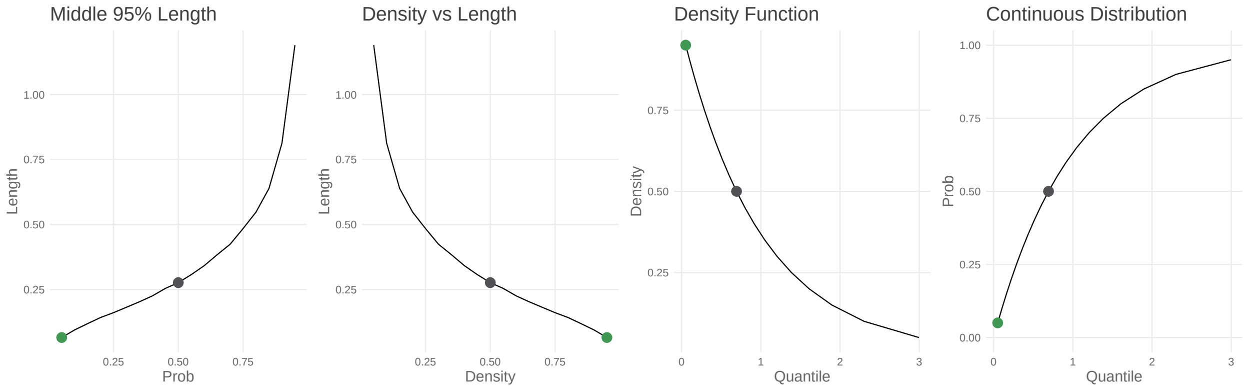
- **Green Dot**: Shortest Length
- **Gray Dot**: Median

Normal Distribution



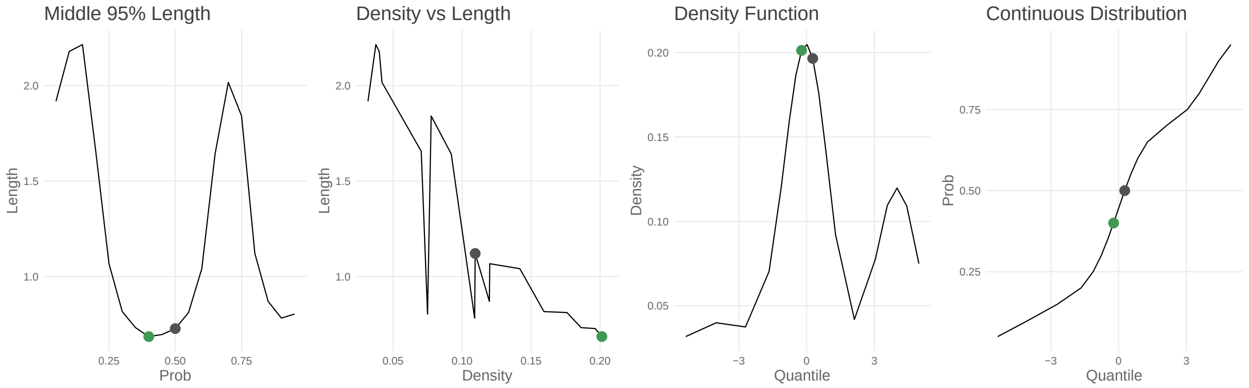
As expected with the normal distributon, the median and shortest point occur very close together. Theoretically, they occur at the exact same point; however, because the random normal distribution was used we should expect a small amount of variance.

Exponential Distribution



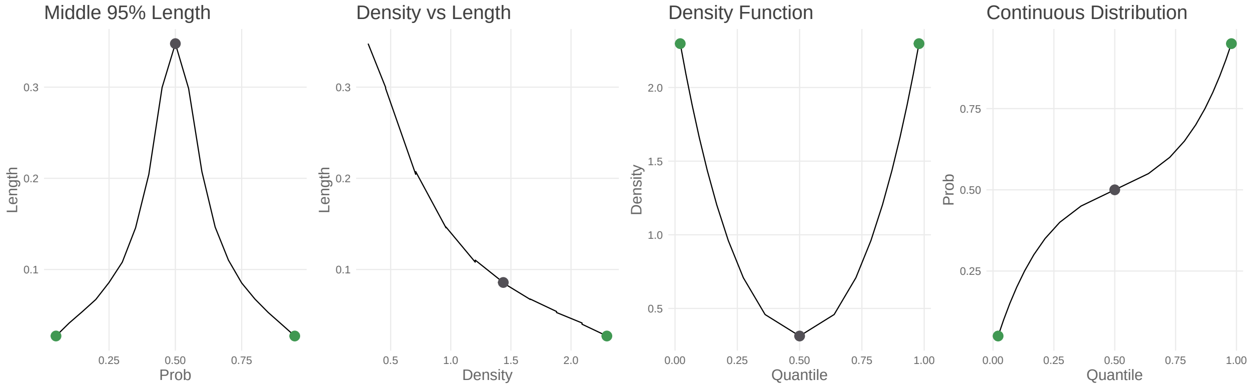
In an exponential distribution values are clustered at the start, which we see with the shortest length occuring at the first value.

Mixed Distribution (F3)



Mixed distribution f3 happens to be similar to the normal, in terms of median and shortest point. This is a great example of why it is important to know the distribution

Mixed Distribution (F4)

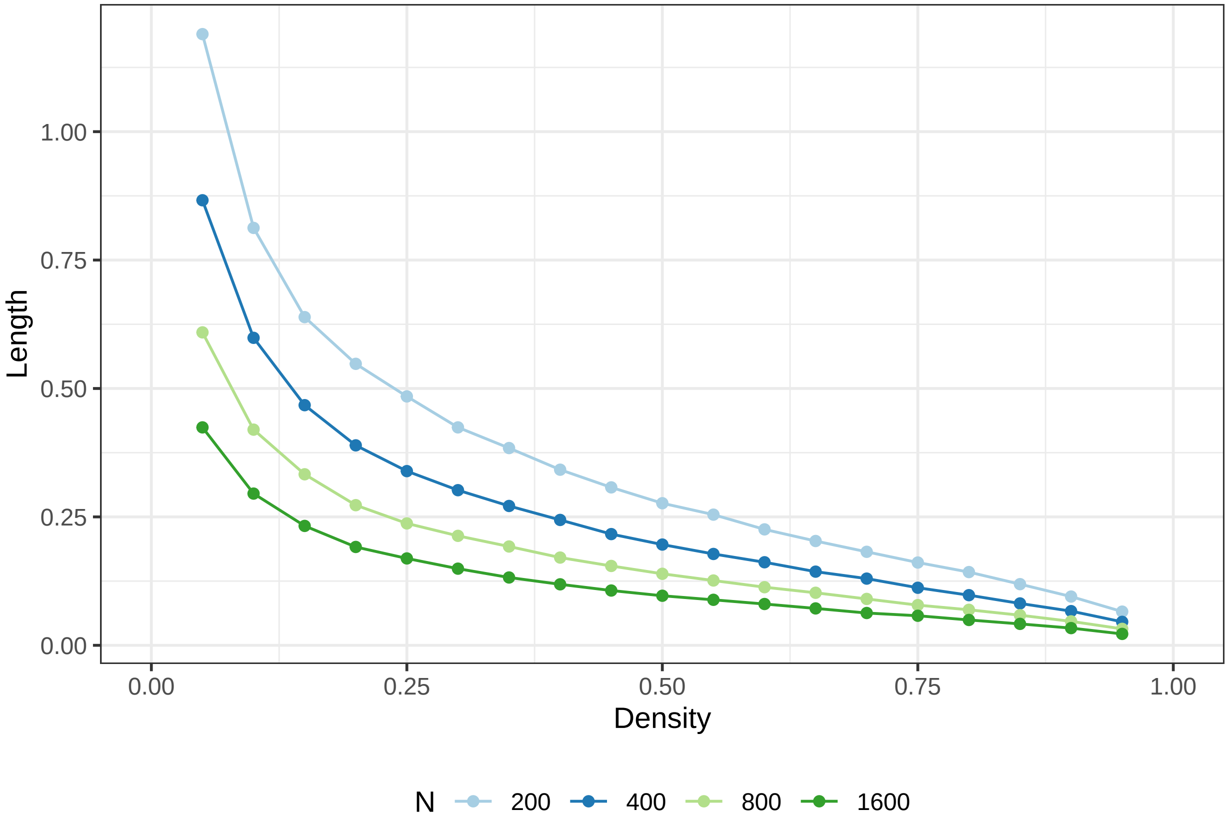


Mixed distriubiton f4 is very close to the opposite of the normal distribution. We can observe that both ends of the distribution (Density Function) are equally dense, resulting in two areas of high relative precision.

The Wrap Up

Hopefully what this illustrated is the importance of correctly identifying or modeling the distribution of a data set. Without it, one cannot quantify the precision of summmary statistics from sample distributions.

With that said, one method of improving precision, is to increase the sample size. Below is a figure that illustrates this for an exponential distribution. As the sample size doubles, the precision is increased a substantial amount. It is important to also notice the diminishing returns as the sample size increases. When using sample distributions, one must strike a balance between precision of estimated summary statistics and avoiding oversampling.


Length
Density
N