Monday, Aug 26

Sampling With Replacement

A variation on simple random sampling is a design where we sample *with replacement* where every possible sample of n elements has the same probability of being selected, but where the elements need not be distinct.

Example: Suppose we have $N = 4$ and $n = 2$ as in previous example. The sampling design for a simple random sampling design when sampling *without replacement* is shown below.

Sample	Probability
$\mathcal{E}_1$, $\mathcal{E}_2$	1/6
$\mathcal{E}_1$, $\mathcal{E}_3$	1/6
$\mathcal{E}_1$, $\mathcal{E}_4$	1/6
$\mathcal{E}_2$, $\mathcal{E}_3$	1/6
$\mathcal{E}_2$, $\mathcal{E}_4$	1/6
$\mathcal{E}_3$, $\mathcal{E}_4$	1/6

But if we sample *with replacement* then the sampling design is as shown below.

Sample	Probability
$\mathcal{E}_1$, $\mathcal{E}_1$	1/16
$\mathcal{E}_1$, $\mathcal{E}_2$	1/16
$\mathcal{E}_1$, $\mathcal{E}_3$	1/16
$\mathcal{E}_1$, $\mathcal{E}_4$	1/16
$\mathcal{E}_2$, $\mathcal{E}_1$	1/16
$\mathcal{E}_2$, $\mathcal{E}_2$	1/16
$\mathcal{E}_2$, $\mathcal{E}_3$	1/16
$\mathcal{E}_2$, $\mathcal{E}_4$	1/16
$\mathcal{E}_3$, $\mathcal{E}_1$	1/16
$\mathcal{E}_3$, $\mathcal{E}_2$	1/16
$\mathcal{E}_3$, $\mathcal{E}_3$	1/16
$\mathcal{E}_3$, $\mathcal{E}_4$	1/16
$\mathcal{E}_4$, $\mathcal{E}_1$	1/16
$\mathcal{E}_4$, $\mathcal{E}_2$	1/16
$\mathcal{E}_4$, $\mathcal{E}_3$	1/16
$\mathcal{E}_4$, $\mathcal{E}_4$	1/16

Sampling *with replacement* changes the properties of the design as well as the sampling distributions of $\bar{y}$ and $\hat{\tau}$.

1. The number of possible samples is N^n, which is larger than the number of possible samples when sampling *without replacement* unless $n = 1$. For example, with a population of $N = 4$ elements and a sample size of $n = 2$, the number of possible samples is 16 when sampling *with replacement* as opposed to 6 when sampling *without replacement*.

2. The inclusion probabilities when sampling *with replacement* are $\pi_i = 1 - (1 - 1/N)^n$ as opposed to $\pi_i = n/N$ when sampling *without replacement*. We will discuss why and how this might be used later in the course.

3. The variance of $\bar{y}$ and $\hat{\tau}$ do not include the finite population correction term so the formulas become

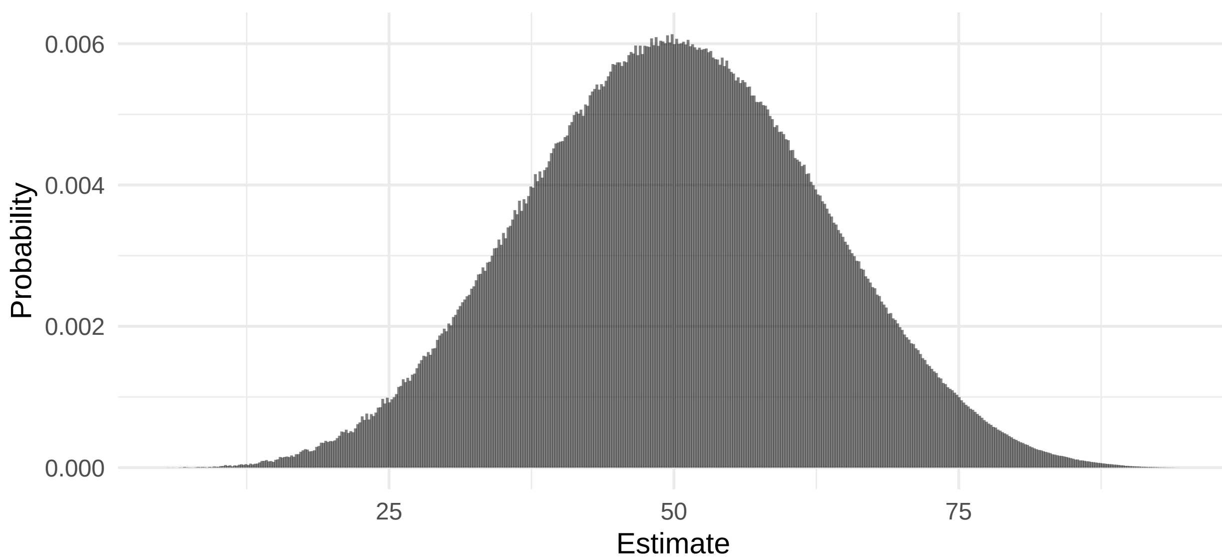
$$V(\bar{y}) = \frac{\sigma^2}{n} \quad \text{and} \quad V(\hat{\tau}) = N^2\frac{\sigma^2}{n}.$$

   So how do the variances of $\bar{y}$ and $\hat{\tau}$ when sampling *with replacement* compare to that when sampling *without replacement*?

Central Limit Theorem

What can we say about the *shape* of the sampling distribution of $\bar{y}$ or $\hat{\tau}$? The **central limit theorem** for simple random sampling states that as n, N, and $N - n$ increase, and assuming some other rather technical but usually applicable conditions, the sampling distribution of $\bar{y}$ "approaches" a normal distribution. This implies the same behavior for the sampling distribution of $\hat{\tau}$.

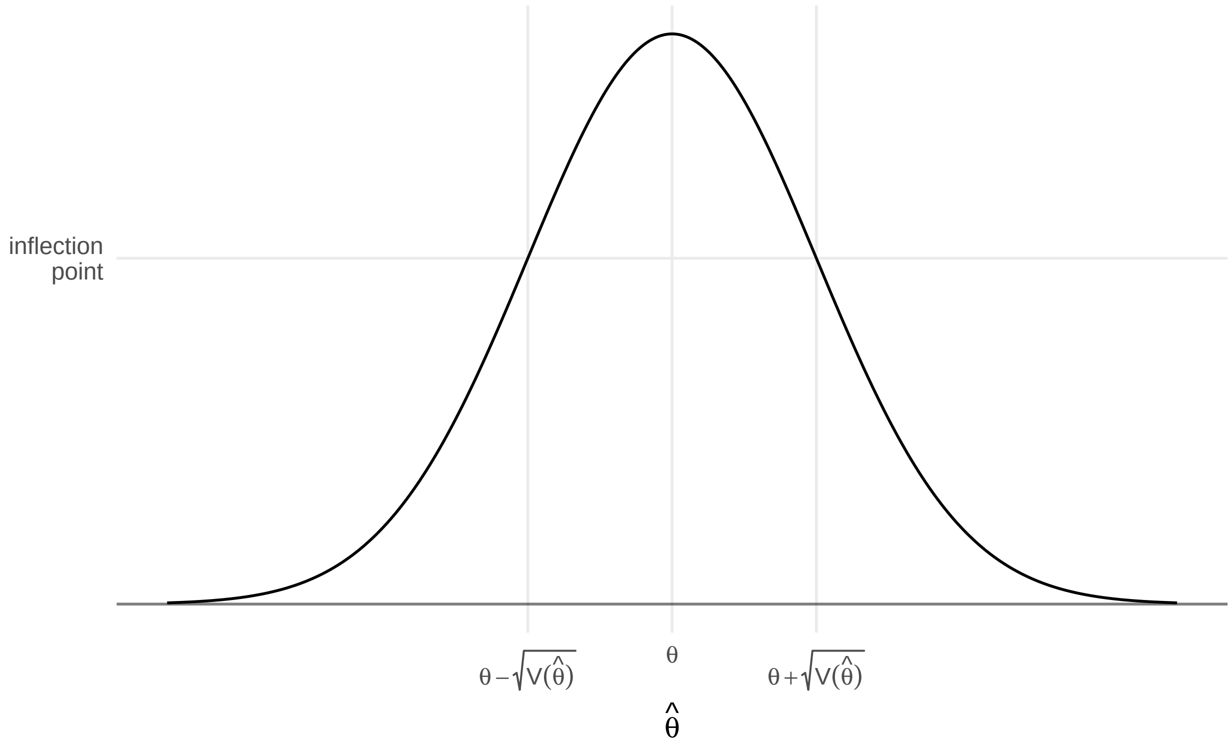
Example: Consider simple random sampling design with a population of $N = 50$ elements and a sample size of $n = 5$. The values of the target variable in the population were selected randomly from the integers from 0 to 100, but the values were "shifted" so that $\mu = 50$. The figure below shows the exact discrete sampling distribution of $\bar{y}$.



Interpreting a Normal Sampling Distribution

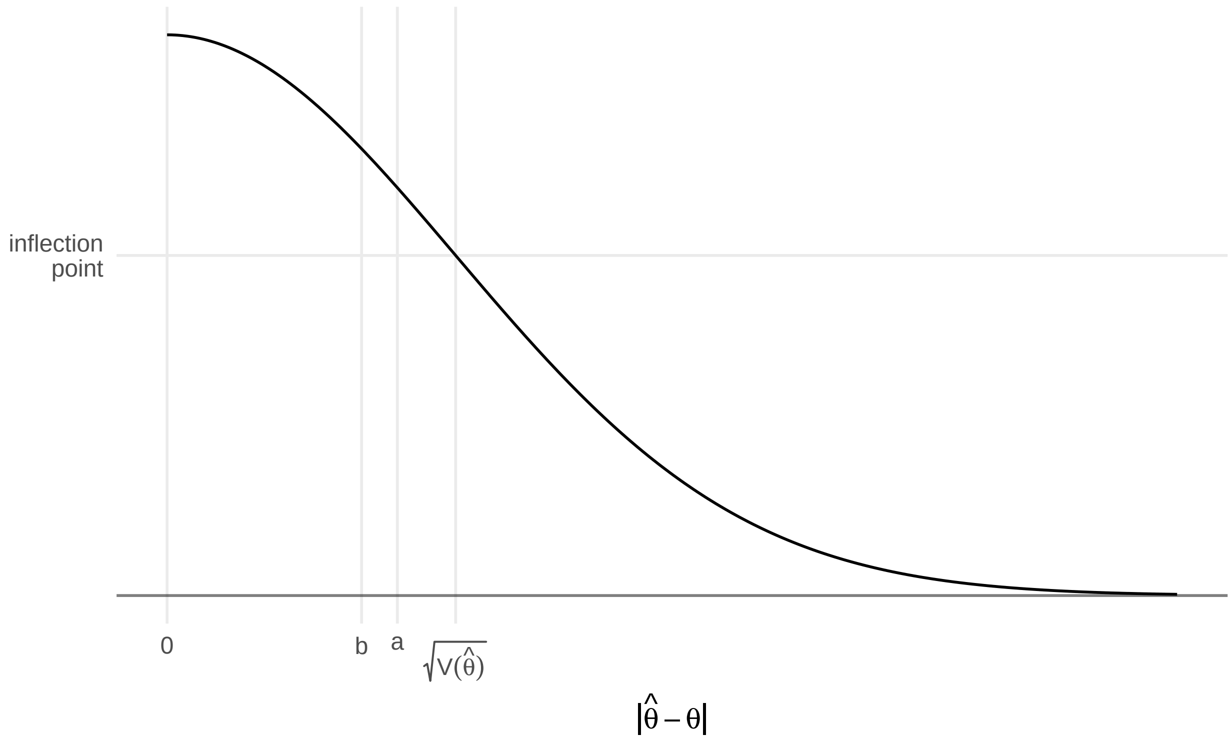
Let θ denote a parameter (e.g., μ or τ) and let $\hat{\theta}$ denote an estimator (e.g., $\bar{y}$ or $\hat{\tau}$). Assume that the sampling distribution of $\hat{\theta}$ is (approximately) normal in shape (by the central limit theorem) with a mean of θ (i.e., the estimator is *unbiased*) and a variance of $\mathrm{V}(\hat{\theta})$. What do we know about this sampling distribution?

1. There is a simple relationship between the mode and inflection points of the probability distribution function and the mean and variance of $\hat{\theta}$.

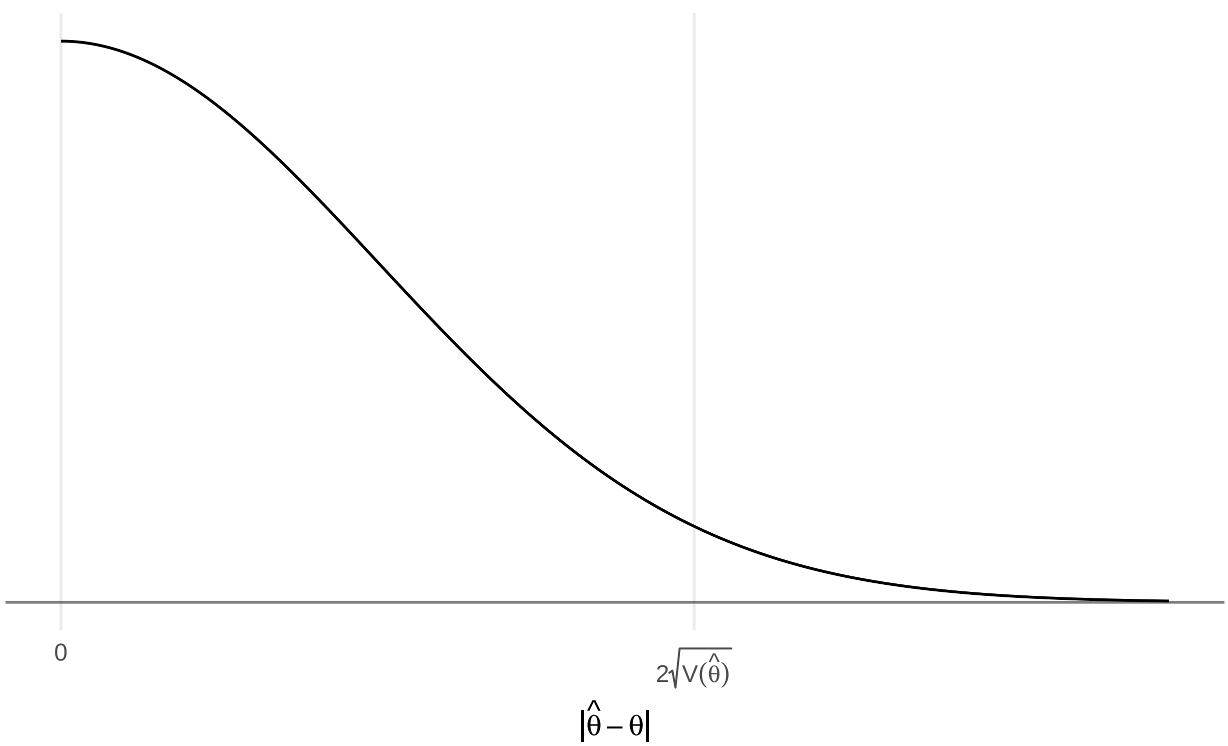


2. The distribution of the *distance* between $\hat{\theta}$ and θ — i.e., $|\hat{\theta} - \theta|$ — is a half-normal distribution with a mean of approximately $0.798\sqrt{\mathrm{V}(\hat{\theta})}$ and a median of approximately $0.674\sqrt{\mathrm{V}(\hat{\theta})}$.[1] These two points are denoted as a and b in the figure below.

[1]The mean is $\sqrt{2\mathrm{V}(\hat{\theta})/\pi}$ and the median is $\sqrt{2\mathrm{V}(\hat{\theta})}\mathrm{erf}^{-1}(0.5)$ where erf^{-1} is the inverse of the error function.



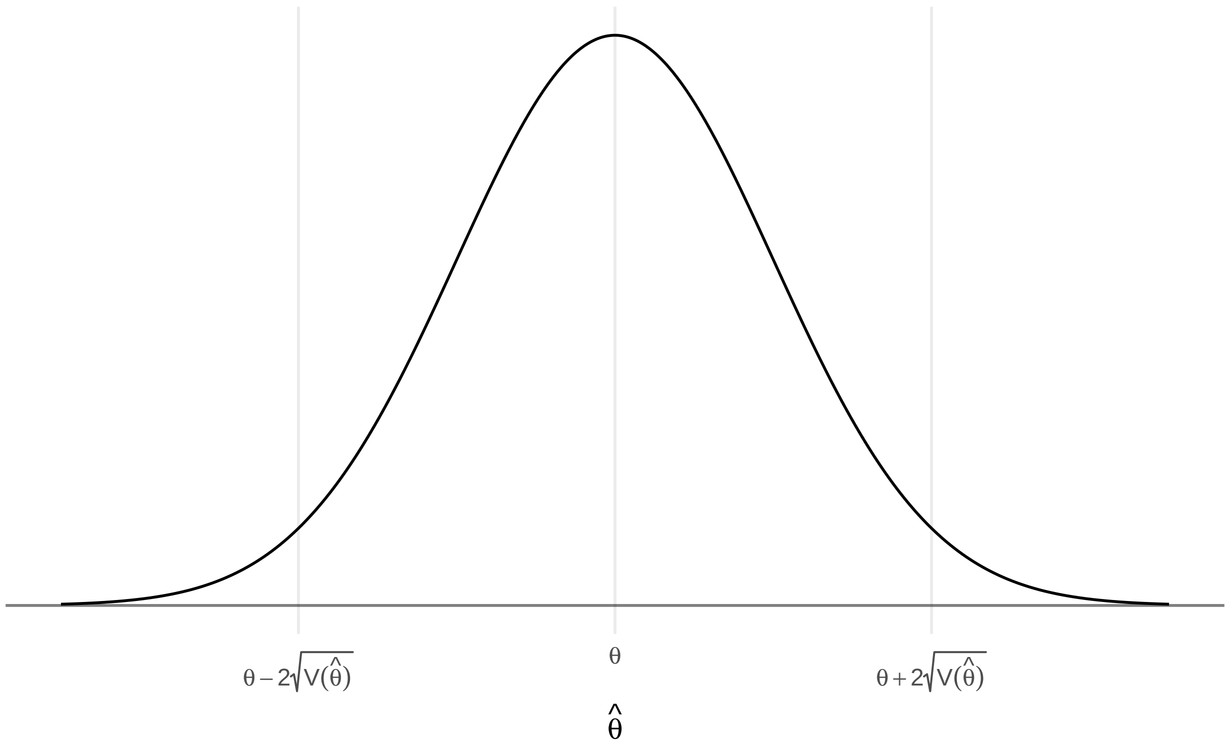
3. The 95th percentile of the *distance* between $\hat{\theta}$ and θ is approximately $2\sqrt{V(\hat{\theta})}$.[2]



This implies that there is approximately a 95% chance that a survey will result in an estimate $\hat{\theta}$ that is within

[2]The actual percentile is closer to $1.96\sqrt{V(\hat{\theta})}$. This can be computed as $\sqrt{2V(\hat{\theta})\text{erf}^{-1}(p/100)}$ where p is the desired percentile (e.g., $p = 95$ for the 95th percentile or $p = 50$ for the 50th percentile and median). Note that any percentile can be computed this way. In survey sampling it is customary to use $2\sqrt{V(\hat{\theta})}$ rather than $1.96\sqrt{V(\hat{\theta})}$ for simplicity.

about $2\sqrt{\mathrm{V}(\hat{\theta})}$ of the parameter θ.



Example: Considered the earlier example with a population of $N = 50$ elements and a simple random sampling design of $n = 5$ elements. The mean and variance for the population are $\mu = 50$ and $\sigma^2 \approx 900$, respectively. What can we conclude about the sampling distribution of $\bar{y}$?

The Standard Error and the Bound on the Error of Estimation

The term $\sqrt{\mathrm{V}(\hat{\theta})}$ is the **standard error** of $\hat{\theta}$. It is simply the standard deviation of $\hat{\theta}$. Under simple random sampling the standard errors of $\bar{y}$ and $\hat{\tau}$ are simply

$$\mathrm{SE}(\bar{y}) = \sqrt{\mathrm{V}(\bar{y})} = \sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}} \quad \text{and} \quad \mathrm{SE}(\hat{\tau}) = \sqrt{\mathrm{V}(\hat{\tau})} = \sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}},$$

respectively. Note that as shown in the previous section many quantities of interest concerning the "accuracy" of an estimator are proportional to the standard error.

The term $2\sqrt{\mathrm{V}(\hat{\theta})}$ is called the **bound on the error of estimation** (also sometimes the **margin of error**). It can be viewed as an kind of upper bound on the distance between $\hat{\theta}$ and θ in the sense that there is about a 95% chance that a survey will not exceed this error. That is

$$P\left(\theta - 2\sqrt{\mathrm{V}(\hat{\theta})} < \hat{\theta} < \theta + 2\sqrt{\mathrm{V}(\hat{\theta})}\right) \approx 0.95.$$

Under simple random sampling the bounds on the error of estimation for $\bar{y}$ and $\hat{\tau}$ are

$$2\sqrt{\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}} \quad \text{and} \quad 2\sqrt{N^2\left(1 - \frac{n}{N}\right)\frac{\sigma^2}{n}},$$

respectively. It is important to note that this result requires that the sampling distribution is normal. In practice it only is true approximately.[3]

Confidence Intervals

The bound on the error of estimation can be used to construct a **confidence interval** which is an "interval estimate" of a parameter (as opposed to a "point estimate") that has a known probability of being correct. This is because

$$P\left(\theta - 2\sqrt{\mathrm{V}(\hat{\theta})} < \hat{\theta} < \theta + 2\sqrt{\mathrm{V}(\hat{\theta})}\right) \approx 0.95$$

implies that

$$P\left(\hat{\theta} - 2\sqrt{\mathrm{V}(\hat{\theta})} < \theta < \hat{\theta} + 2\sqrt{\mathrm{V}(\hat{\theta})}\right) \approx 0.95.$$

The confidence interval can be written as

$$\hat{\theta} \pm 2\sqrt{\mathrm{V}(\hat{\theta})} \Leftrightarrow \left(\hat{\theta} - 2\sqrt{\mathrm{V}(\hat{\theta})}, \hat{\theta} + 2\sqrt{\mathrm{V}(\hat{\theta})}\right).$$

The probability of 95% is the **confidence level** of the confidence interval. It represents the expected percent of confidence intervals that would correctly estimate the parameter, and the probability that a survey will produce an estimate with an error less than the bound on the error of estimation.

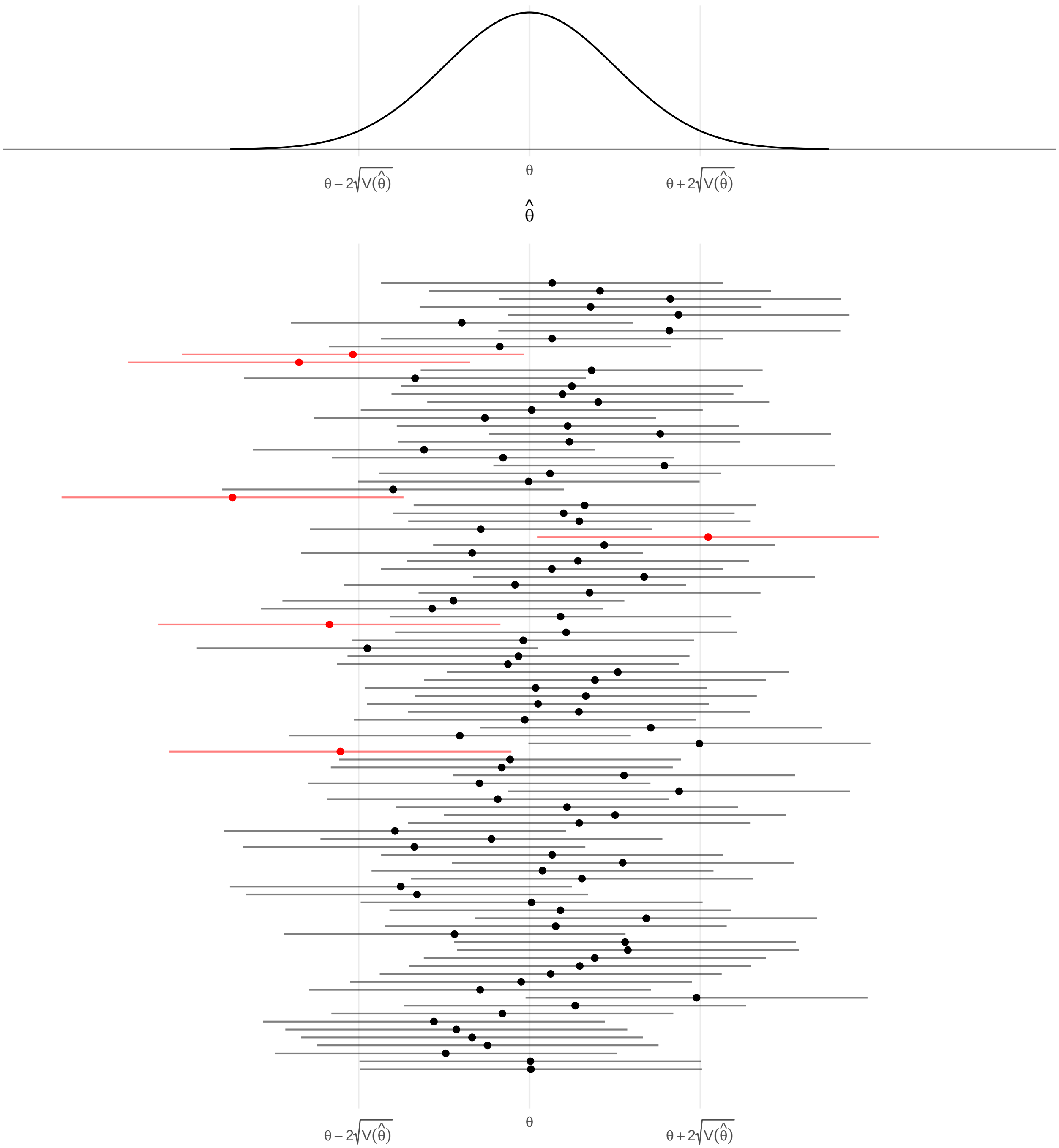
[3] A more general statement can be made that is true for *any* type of sampling distribution for an unbiased estimator. There is a result called Chebyshev's inequality that implies that

$$P\left(\theta - k\sqrt{V(\hat{\theta})} < \hat{\theta} < \theta + k\sqrt{V(\hat{\theta})}\right) > \frac{1}{k^2}$$

for any $k > 0$. That is, the probability that $\hat{\theta}$ is within k standard errors of θ is at least $1/k^2$. If we let $k = \delta/\sqrt{V(\hat{\theta})}$ then we can say that

$$P(\theta - \delta < \hat{\theta} < \theta + \delta) > \frac{V(\hat{\theta})}{\delta^2}.$$

Thus the lower bound on the probability that $\hat{\theta}$ is within δ of θ is proportion to $V(\hat{\theta})$. So regardless of the shape of the sampling distribution the variability of the sampling distribution plays an important role in how close $\hat{\theta}$ might be to θ.



Example: Consider the previous example. If we obtain a sample and compute sample mean of $\bar{y} = 71.64$, what is the *confidence interval* for estimating μ? What is the confidence interval for estimating τ?

Variance Estimation

We do not typically know σ^2 (i.e., the variance of the target variable for the population), although we might use an educated guess if necessary (more on that later). But we can **estimate** it after a survey has been conducted, and use that to estimate the variance of an estimator. An unbiased estimator is the *sample variance*

$$s^2 = \frac{1}{n-1}\sum_{i \in \mathcal{S}}(y_i - \bar{y})^2.$$

Then the estimators of the variance of $\bar{y}$ and $\hat{\tau}$ are

$$\widehat{\mathrm{V}(\bar{y})} = \left(1 - \frac{n}{N}\right)\frac{s^2}{n} \quad \text{and} \quad \widehat{\mathrm{V}(\hat{\tau})} = N^2\left(1 - \frac{n}{N}\right)\frac{s^2}{n}.$$

The problem of *estimating the variance of an estimator* is called **variance estimation** in survey sampling.

Example: The sample from the previous example yields a sample variance of $s^2 = 616.7$. The sample is 78.84, 29.84, 90.84, 88.84, and 69.84. How would we use this to compute (a) the bound on the error of estimation for estimating μ with $\bar{y}$ and (b) a confidence interval for estimating μ?