HANDOUT #10: SAMPLING DISTRIBUTIONS

1. Sampling Distributions
   (a) Definition
   (b) Expected Value and Variance of Statistic
   (c) MSE and Bias

2. Methods for Determining Sampling Distribution:
   (a) Enumeration of All Possible Outcomes: Age of Coin Example
   (b) Theoretical Sampling Distribution: Mathematical Derivation
   (c) Simulation Results: Simulate Realizations from Specified Distributions
   (d) Asymptotic Results: Various Central Limit Theorems
   (e) Bootstrap Methods: Resampling From An Observed Sample

3. Sampling Distribution of
   (a) Sample Mean - $\bar{Y}$
   (b) Sample Median - $\widehat{Q}(.5) = \tilde{Y}$
   (c) Sample Standard Deviation - S
   (d) Sample Proportion - $\hat{p}$
   (e) Sample Quantiles - $\widehat{Q}(u)$
   (f) Maximum Likelihood Estimator - $\widehat{\theta}$

Sampling Distributions

Definition of Sampling Distribution of a Statistic:

After obtaining an estimator of a population parameter ($\mu, \sigma, Q(u)$) or a parameter associated with a particular family of distributions (β in an exponential family or β and γ in an Weibull family), the **Sampling Distribution of the estimator** is distribution of the estimator as its possible realizations vary across all possible samples that may arise from a given population.

For example, let θ be a parameter for a population or for a family of distributions. Let $Y_1, \ldots, Y_n$ be iid random variables with cdf F completely unspecified or $F(\cdot, \theta)$, where θ is a vector of unknown parameters. Let $\hat{\theta} = \hat{\boldsymbol{\theta}}(Y_1, \ldots, Y_n)$ be an estimator of θ based on the observed data. We want to assess how well does $\hat{\theta}$ estimate θ. Some measures of this assessment are given here:

- Concentration of the values of $\hat{\theta}$ about θ:

  What is the chance that $\hat{\theta}$ will be close to θ? That is,

  Compute $P[|\hat{\theta} - \theta| < \epsilon]$ for small values of ϵ.

- On the average does $\hat{\theta}$ equal θ?

  Compute the **Bias** of using $\hat{\theta}$ as an estimator θ:

  Bias$=E[\hat{\theta}] - \theta$.

  If Bias=0, we state that $\hat{\theta}$ is an **unbiased estimator** of θ.

- On the average is $\hat{\theta}$ close to θ?

  Compute the average squared distance from $\hat{\theta}$ to θ, the **Mean Squared Error**:

$$MSE(\hat{\theta}) = E[(\hat{\theta} - \theta)^2] = Var(\hat{\theta}) + [Bias]^2$$

In order to calculate the above quantities, we need to know the distribution of $\hat{\theta}$ over all possible samples from the population. This would be nearly impossible for a large population but we can envision the procedure as follows:

- Take M samples of size n from the population

  Sample 1: $X_{11}, \ldots, X_{1n}$ then compute $\hat{\theta}_1$
  Sample 2: $X_{21}, \ldots, X_{2n}$ then compute $\hat{\theta}_2$
  $\ldots$
  Sample M: $X_{M1}, \ldots, X_{Mn}$ then compute $\hat{\theta}_M$

- Estimate the distribution of $\hat{\theta}$ using the M realizations of $\hat{\theta}$: $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$.

- We could then estimate the cdf of $\hat{\theta}$, its mean, variance, bias, MSE, etc.

In nearly all situations, this procedure is impossible because of cost, time, or the mere impossibility of every being able to take enough repeated samples from a fixed population. We can overcome this problem in a number of ways using one or a combination of the following procedures:

I. Mathematical Derivations: Consider the following examples

1. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then the distribution of $\hat{\mu} = \bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$
2. If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then the distribution of $\hat{\sigma} = S$ can be obtained from the result

   $(n-1)S^2/\sigma^2$ is distributed chi-square with $df = n - 1$.
3. If $T_1, \ldots, T_n$ are iid $Exp(\beta)$, then then the distribution of $n\hat{\beta} = \sum_{i=1}^n T_i$ is distributed $Gamma(n, \beta)$
4. If $Y_1, \ldots, Y_n$ are iid Bernoulli with unknown value p, then the distribution of $\hat{p} = \bar{Y}$ is obtained from the result that $n\hat{p} = \sum_{i=1}^{n} Y_i$ is distributed $B(n, p)$

II. Asymptotic theory (large n) can also be applied in many situations.

1. Various versions of the central limit gives approximations to the sampling distributions of $\bar{Y}$, $\hat{p}$, $\hat{\sigma}$, $\hat{Q}(.5)$, etc.
2. If $\hat{\theta}$ is a MLE of θ then there are versions of the central limit theorem which describe the sampling distribution of $\hat{\theta}$.
3. The extreme value distributions can be used to approximate the sampling distributions of the sample minimum $Y_{(1)}$ and maximum $Y_{(n)}$.

III. Simulation studies provide some insight to the sampling distribution but are limited in that we must specify the population distribution exactly in order to conduct the simulation. Consider the following example:

Suppose we wanted to determine the sampling distribution of the estimators of (θ_1, θ_2) when sampling from a population having a Cauchy distribution. We can simulate observations from a Cauchy distribution using R but first we must design the simulation study.

1. How many different values of the sample size n will be needed?
2. How many different values of the location parameter θ_1 will be needed?
3. Which values of θ_1 should be selected?
4. How many different values of the scale parameter θ_2 will be needed?
5. Which values of θ_2 should be selected?
6. How many replications of the simulation are needed for each choice of (n, θ_1, θ_2)
7. How can we infer the sampling distribution of $(\hat{\theta}_1, \hat{\theta}_2)$ for values of (n, θ_1, θ_2) not run in the simulation study.

IV. Suppose we have a random sample of n units from a population with n of a modest size. We want to determine the sampling distribution of the sample median. The population distribution is completely unknown hence a simulation study can not be utilized. The sample size is too small to have much confidence in applying the central limit theorem for the median. A possible method for obtaining an approximation to the sampling distribution of the statistics is to use a resampling procedure. This involves taking numerous samples of size n (with replacement) from the actual observed data and computing the value of the statistic from each of these samples. One such procedure is called the **bootstrap sampling procedure**.

We will discuss each of these procedures briefly. Greater detail will be provided in other statistics courses: STAT 610-611, 630, and 689 courses on resampling methods.

There are many ways to demonstration the sampling distributions of various sample statistics. We will first have a physical demonstration involving the age of pennies.

Physical Demonstration of Sampling Distribution of $\bar{X}$

The following example will demonstrate the central limit theorem through the use of the ages of 500 pennies.

I have a collection of 500 pennies with X equal to the Age of penny: $X = 2012$ - (Date on penny).

What general shape do you think a histogram of age, X, would have? Why?

Stem and Leaf Plot for the Ages of 500 Pennies

37 - 0	|	1111111111111111111111111111111111111
85 - 0	|	2222222222222222222222222222222222222222222222222223333333333333333333333333333333333
38 - 0	|	44444444444444444444445555555555555555
44 - 0	|	66666666666666666666667777777777777777777777
37 - 0	|	8888888888888888888888888899999999999
18 - 1	|	000000000000001111
38 - 1	|	22222222222222222333333333333333333333
46 - 1	|	4444444444444444444444555555555555555555555555
30 - 1	|	666666666667777777777777777777
20 - 1	|	88888888888888899999
21 - 2	|	000000000000000111111
18 - 2	|	222222233333333333
18 - 2	|	444444445555555555
17 - 2	|	66667777777777777
9 - 2	|	888889999
1 - 3	|	1
8 - 3	|	33333333
8 - 3	|	44444555
4 - 3	|	6777
2 - 3	|	88
1 - 4	|	0

This is a very right skewed distribution with five-number summary of the distribution:

$$\text{Minimum } = 1, \qquad Q_1 = 5, \qquad M = Q_2 = 10, \qquad Q_3, = 18, \qquad \text{Maximum } = 40$$

The mean and standard deviation for this population of all 500 pennies in the population were computed to be:

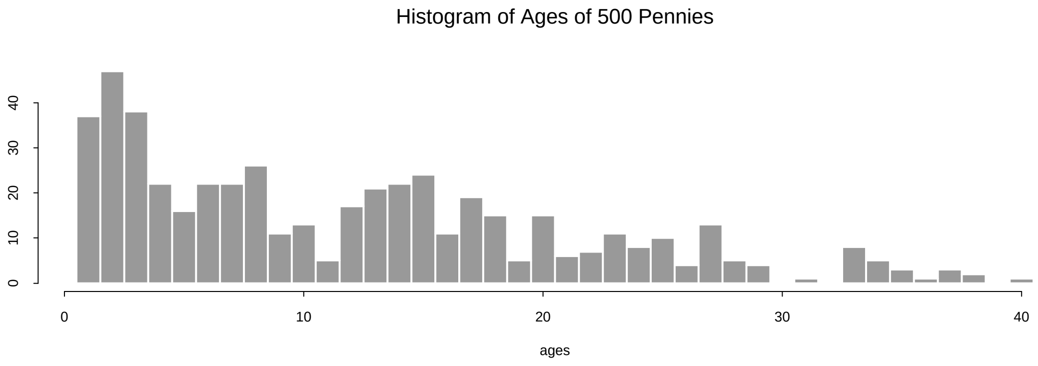
$$\mu_X = 12.038 \text{ and } \sigma_X = 9.281$$

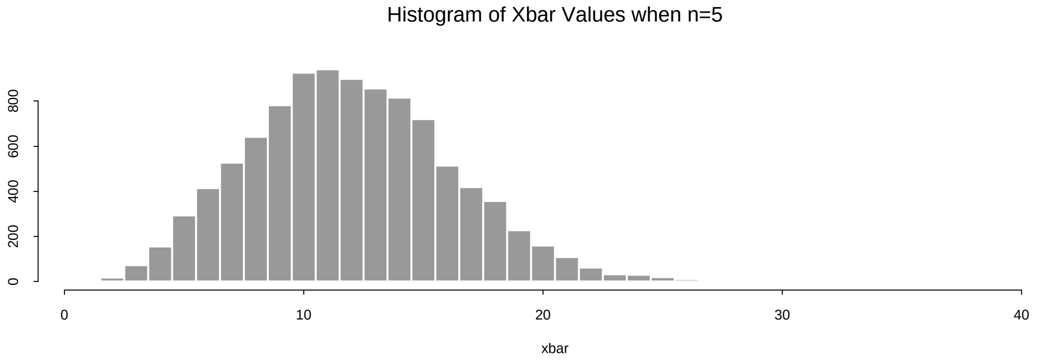
Sampling Distribution of Sample Mean, $\bar{X}$: To obtain the histogram of the sample means for *all* possible samples of a selected size n would be extremely difficult because of the enormous number of possible samples of size n:

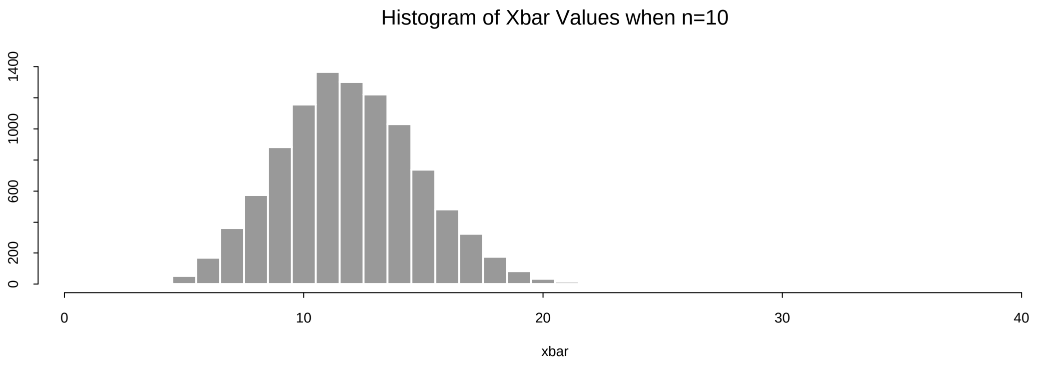
n	Number of Possible Samples of Size n: $\binom{500}{n}$
5	255,244,687,600
10	2.458×10^{20}
20	2.667×10^{35}

Because of the enormous number of possible samples of size n needed to completely describe the sampling distribution of $\bar{X}$, approximations to the sampling distribution of $\bar{X}$ were generated by taking 10,000 random samples of size 5, 10, and 20 from the population of 500 coins.

Compare the shapes and spreads of the sampling distributions of $\bar{X}$ with the distribution of the original population.







Notice how the shape of the sampling distribution of $\bar{X}$ tends towards a symmetric distribution (normal distribution) as n increases from n=1 to 5 to 10 to 20.

Sampling Distribution of $\bar{X}$, $\hat{Q}(.5)$, S, $\hat{p}$

Let $X_1, \ldots, X_n$ be iid random variables with cdf F having $\mu, \sigma < \infty$ for its mean and standard deviation and $\mu_3 < \infty$ and $\mu_4 < \infty$ as its 3rd and 4th central moments.

Sampling Distribution of the Sample Mean: $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- For all n : If we sample with replacement,

  $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \frac{\sigma^2}{n}$

- For all n : If we sample without replacement,

  $E[\bar{X}] = \mu$ and $Var[\bar{X}] = \left(\frac{N-n}{N-1}\right)\frac{\sigma^2}{n} < \frac{\sigma^2}{n}$,

  where N is the population size and $\left(\frac{N-n}{N-1}\right)$ is called the finite population correction factor (fpcf).

- For all values of n : If F is $N(\mu, \sigma^2)$, then the distribution of $\bar{X}$ is $N(\mu, \frac{\sigma^2}{n})$.

- The Central Limit Theorem yields for large n : The distribution of $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ converges to a $N(0,1)$ as $n \to \infty$, provided $\sigma < \infty$. That is, for large n, the distribution of $\bar{X}$ is approximately $N(\mu, \frac{\sigma^2}{n})$.

- The accuracy of the approximate distribution of $\bar{X}$ depends on the size of n and the shape of F. The larger the value of n, the better the approximation. For a fixed value of n, the more symmetric with normal-like tails that F is, the greater the accuracy of the approximation.

- If the data is correlated, then $\bar{X}$ is still an unbiased estimator of μ but the variance of $\bar{X}$ may be larger or smaller than $\frac{\sigma^2}{n}$ depending on the type of correlation in the data.

- For example, suppose $corr(X_i, X_j) = \rho$ for all pairs $i \neq j$, (equi-correlated), where $\frac{-1}{n-1} < \rho < 1$. Then,

$$E(\bar{X}) = \mu \qquad \text{but} \qquad Var(\bar{X}) = \frac{\sigma^2}{n}[1 + (n-1)\rho] \Rightarrow \begin{cases} Var(\bar{X}) < \frac{\sigma^2}{n} & \text{if } \rho < 0 \\ Var(\bar{X}) > \frac{\sigma^2}{n} & \text{if } \rho > 0 \end{cases}$$

To verify the above, recall the following result for constants $c_1, c_2, \ldots, c_n$ and random variables $X_1, X_2, \ldots, X_n$:

$$Var\left(\sum_{i=1}^{n} c_i X_i\right) = \sum_{i=1}^{n} c_i^2 Var(X_i) + \sum_{i \neq j} c_i c_j Cov(X_i, X_j)$$

- Second example: Suppose that the $X_i s$ follow an AR(1) process: $X_t = \theta + \rho X_{t-1} + e_t$, where $X_t s$ are stationary and $e_t s$ are iid with $E[e_t] = 0, Var(e_t) = \sigma_e^2$, with $-1 < \rho < 1$. Then, we have

  $\mu_t = E[X_t] = \theta/(1-\rho), \quad \sigma_t^2 = Var(X_t) = \sigma_e^2/(1-\rho^2), \quad Corr(X_i, X_j) = \rho^{|i-j|}$ for $i \neq j$,

  The values of μ_t and σ_t are constant and do not depend on t. The amount of correlation decreases as time or distance between two units increases.

  We have $E(\bar{X}) = \mu$ but

$$Var(\bar{X}) = \frac{\sigma^2}{n}\left[1 + \frac{2}{n}\left(\frac{\rho}{1-\rho}\right)\left(n + \frac{1-\rho^n}{1-\rho}\right)\right] \approx \frac{\sigma^2}{n}\left[\frac{1+\rho}{1-\rho}\right] \Rightarrow$$

$$Var(\bar{X}) < \tfrac{\sigma^2}{n} \quad \text{if} \quad \rho < 0$$

$$Var(\bar{X}) > \tfrac{\sigma^2}{n} \quad \text{if} \quad \rho > 0$$

- For *iid* data, $\widehat{SE}(\bar{X}) = S/\sqrt{n}$

  This estimate would underestimate $SE(\bar{X})$ if $X_i s$ are AR(1) with $\rho > 0$ which would result in a confidence interval for μ having coverage probability much less than the stated level, e.g., coverage probability of 80% for a stated 95% C. I.

  A better estimator would be $\widehat{SE}(\bar{X}) = \frac{S}{\sqrt{n}}\sqrt{\frac{1+\hat{\rho}}{1-\hat{\rho}}}$

  For example, consider the ozone data and suppose that an AR(1) model is adequate. Then, we have the following results

City	n	S	$\hat{\rho}$	$S/\sqrt{n}$	$\widehat{SE}(\bar{X})$	% $S/\sqrt{n}$ is less than $\widehat{SE}(\bar{X})$
Yonkers	148	28.11	0.4342	2.31	3.68	59.3%
Stamford	136	52.11	0.3342	4.47	6.33	41.6%

Sampling Distribution of the Sample Median: $\tilde{\mu} = \hat{Q}(.5)$

Let $\hat{Q}(.5)$ be the sample median given by

$$\hat{Q}(.5) = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ .5\left(X_{(\frac{n}{2})} + X_{(\frac{n+2}{2})}\right) & \text{if } n \text{ is even} \end{cases}$$

- The moments of $\hat{Q}(.5)$ depend on F and the joint distribution of the order statistics: For n odd

$$m_k = E\left[\left(\hat{Q}(.5)\right)^k\right] = E\left[X^k_{(\frac{n+1}{2})}\right] = n\binom{n-1}{\frac{n-1}{2}} \int_{-\infty}^{\infty} x^k \left[F(x)[1-F(x)]\right]^{\frac{n-1}{2}} f(x)dx$$

- The mean of $\hat{Q}(.5)$ is m_1 and the variance is $m_2 - m_1^2$,.

  Both of which depend on the population cdf F

  Similar results are obtained for n even.

- For large n : The distribution of $\hat{Q}(.5)$ is approximately $N\left(Q(.5), \frac{(.5)^2}{n(f(Q(.5))^2}\right)$, thus, the asymptotic mean and standard deviation for the sample median $\hat{Q}(.5)$ are given by

$$\mu_A = Q(.5) \text{ and } \sigma_A = \frac{.5}{f(Q(.5))\sqrt{n}} = \frac{.5/f(Q(.5))}{\sqrt{n}}.$$

  That is, $\hat{Q}(.5)$ is asymptotically unbiased but for small n it would be a biased estimator of Q(.5) in most situations.

- For estimating the location parameter θ of a symmetric distribution with $\sigma < \infty$,

  should we use $\bar{X}$ or $\hat{Q}(.5)$?

Definition: The Asymptotic Relative Efficiency (ARE) of $\widehat{Q}(.5)$ to $\bar{X}$ as an estimator of θ is given by

$$ARE(\hat{Q}(.5), \bar{X}) = \frac{\text{asymptotic variance of } \bar{X}}{\text{asymptotic variance of } \widehat{Q}(.5)} = \frac{\sigma^2/n}{\left(\frac{.5/f(Q(.5))}{\sqrt{n}}\right)^2} = 4\sigma^2 f^2(Q(.5))$$

When $ARE < 1$, the Sample Mean= $\bar{X}$ is a more efficient estimator of θ than the Sample Median=$\widehat{Q}(.5)$

1. F has a $N(\theta, \sigma^2)$, then $f(x) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{\frac{(x-\theta)^2}{2\sigma^2}}$, $Q(.5) = \theta$ and $f(Q(.5)) = f(\theta) = \frac{1}{\sigma\sqrt{2\pi}}$

   ARE $= 4\sigma^2 f^2(Q(.5)) = 4\sigma^2 \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^2 = \frac{2}{\pi} \approx .64 < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

2. F has a uniform on $(\theta - a, \theta + a)$, then $f(x) = \frac{1}{2a} I(\theta - a < x < \theta + a)$, $Q(.5) = \theta$, $\sigma^2 = \frac{a^2}{3}$

   $f(Q(.5)) = f(\theta) = \frac{1}{2a}$ which yields

   ARE $= 4\sigma^2 f^2(Q(.5)) = 4\frac{a^2}{3}\frac{1}{4a^2} = \frac{1}{3} < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

3. F is a logistic(θ_1, θ_2), then $f(x) = \frac{e^{-(x-\theta_1)/\theta_2}}{\theta_2(1+e^{-(x-\theta_1)/\theta_2})^2}$, $Q(.5) = \theta_1$, $\sigma^2 = \frac{\pi^2\theta_2^2}{3}$, and

   $f(Q(.5)) = f(\theta_1) = \frac{1}{4\theta_2}$ which yields

   ARE $= 4\sigma^2 f^2(Q(.5)) = 4\frac{\pi^2\theta_2^2}{3}\left(\frac{1}{4\theta_2}\right)^2 = \frac{\pi^2}{12} \approx .82 < 1$

- Sample Mean is more efficient estimator of θ than is the Sample Median

4. F has a shifted t distribution with two parameters: shift θ_1 and shape $\theta_2 = df > 2$,

   then $Q(.5) = \theta_1$, $\sigma^2 = \frac{\theta_2}{\theta_2 - 2}$, and $f(Q(.5)) = f(\theta_1) = \frac{\Gamma(\frac{\theta_2+1}{2})}{\Gamma(\frac{\theta_2}{2})\sqrt{\pi\theta_2}}$ which yields

$$\text{ARE}{=}4\sigma^2 f^2(Q(.5)) = 4\left(\frac{\theta_2}{\theta_2 - 2}\right)\left[\frac{\Gamma(\frac{\theta_2+1}{2})}{\Gamma(\frac{\theta_2}{2})\sqrt{\pi\theta_2}}\right]^2$$

θ_2	3	4	5	8	∞
ARE	1.62	1.12	0.96	0.8	0.64

For $2 < df \leq 4$, the Median is a more efficient estimator of the shift parameter, θ_1 than the Mean.

For $5 < df$, the Mean is a more efficient estimator of the shift parameter, θ_1 than the Median.

Thus, for a heavy-tailed, symmetric distribution, the Sample Median is a more efficient (less variable) estimator of the location parameter in comparison to the Sample Mean.

Sampling Distribution of the Sample Standard Deviation: S

Let $S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ be the sample standard deviation.

- For all n : $E[S^2] = \sigma^2$
- For all n : $E[S] \neq \sigma$. Why?
- The $Var[S]$ depends on F
- For all n : If F is $N(\mu, \sigma^2)$, then the distribution of $\frac{(n-1)S^2}{\sigma^2}$ is chi-square with $df = n-1$
- For all n : If F is $N(\mu, \sigma^2)$, then
  - $E[S] = \left[\sqrt{\frac{2}{n-1}}\left(\frac{\Gamma(\frac{n}{2})}{\Gamma(\frac{n-1}{2})}\right)\right]\sigma = [c_n]\sigma, \quad \text{with} \quad c_n\sigma \rightarrow \sigma \quad \text{as} \quad n \rightarrow \infty$
  - $Var[S] = \sigma^2(1 - c_n^2) \quad \text{with} \quad \sigma^2(1-c_n^2) \rightarrow 0 \quad \text{as} \quad n \rightarrow \infty$
- The bias in using S as an estimator of σ when the data is from a normally distributed population: Bias $= E[S] - \sigma = (c_n - 1)\sigma$

n	2	3	4	5	10	25	100	250
Bias	-.202σ	-.114σ	-.079σ	-.060σ	-.027σ	-.0104σ	-.00252σ	-.00100σ

- For large n : The distribution of S is approximately $N\left(\sigma, \frac{\mu_4 - \sigma^4}{4n\sigma^2}\right)$, thus, the asymptotic mean and standard deviation for the sample standard deviation S are given by

$$\mu_A = \sigma \quad \text{and} \quad \sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}}.$$

- How accurate is this approximation? As is true for all asymptotic results, it depends on n and the population distribution F. A simulation study will demonstrate the accuracy of using the normal approximation in place of the true sampling distribution of $\bar{X}$, $\hat{Q}(.5)$, and S.
- How does either spatial or temporal correlation affect S as an estimator of σ?

  Let $Y_1, Y_2, \ldots, Y_n$ be stationary random variables with mean μ, variance σ^2, and covariance function $R(k) = E[(Y_t - \mu)(Y_{t+k} - \mu)]$. Then we have the following results.

  - $E[\bar{Y}] = \mu$ $\qquad$ $E[S^2] = \sigma^2 = R(0)$
  - $Var(\bar{Y}) = \frac{1}{n}\sum_{i=-n}^{n}\left(1 - \frac{i}{n}\right)R(i)$ $\qquad$ $Var(S^2) = \frac{2}{n}\sum_{i=-n}^{n}\left(1 - \frac{i}{n}\right)R^2(i)$

Sampling Distribution of the Sample Quantiles: $\hat{Q}(u)$

Let $\hat{Q}(u)$ be the sample quantile for values of u not too close to 0 or 1.

- The quantities $E[\widehat{Q}(u)]$ and $Var[\widehat{Q}(u)]$ depend on F through the distribution of the order statistics. Thus, we can compute these values using our smoothed definition of the sample quantile function:

  Let $\frac{1}{2n} \leq u \leq 1 - \frac{1}{2n}$ with $nu + .5 = k + r$ where

  $k = 1, \cdots, n-1$ and $0 < r < 1$ then we define

  $\widehat{Q}(u) = Y_{(k)} + r[Y_{(k+1)} - Y_{(k)}]$

  The pdf of the kth order statistic, $Y_{(k)}$ is given by

  $$f_{(k)}(y) = \frac{n!}{(k-1)!(n-k)!} [F(y)]^{k-1} [1 - F(y)]^{n-k} f(y)$$

  From which we can compute the mean of $Y_{(k)}$ which then yields

  $$E\left[\widehat{Q}(u)\right] = (1-r)E\left[Y_{(k)}\right] + E\left[Y_{(k+1)}\right]$$

  Similarly, using the joint pdf of two order statistics, we can obtain the covariance between any two order statistics. The $Var[\widehat{Q}(u)]$ can then be computed using the expressions for the variances and covariance of the order statistics.

- For large n : The distribution of $\hat{Q}(u)$ is approximately $N\left(Q(u), \frac{u(1-u)}{n(f(Q(u))^2}\right)$, thus, the asymptotic mean and standard deviation for the sample quantile $\hat{Q}(u)$ are given by

  $$\mu_A = Q(u) \qquad \text{and} \qquad \sigma_A = \frac{\sqrt{u(1-u)}}{f(Q(u))\sqrt{n}} = \frac{\sqrt{u(1-u)}/f(Q(u))}{\sqrt{n}}.$$

  For a $N(\mu, \sigma^2)$ distribution, $f(Q(u)) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(Q(u)-\mu)^2/2\sigma^2}$

  For the median, $Q(.5) = \mu$ therefore $f(Q(.5)) = \frac{1}{\sigma\sqrt{2\pi}}$ thus

  $\sigma_A = \frac{\sqrt{.5(1-.5)}}{f(Q(.5))\sqrt{n}} = \frac{\sigma\sqrt{\frac{\pi}{2}}}{\sqrt{n}}$. This formula is only for the normal distribution.

  In practice, we would need to estimate $f(Q(u)))$ in order to obtain a value for σ_A.

Sampling Distribution of the Sample Minimum and Maximum

Let $X_1, \ldots, X_n$ iid with cdf F.

Define

$M_n = X_{(n)} = max[X_1, \ldots, X_n]$ as the sample maximum

$m_n = X_{(1)} = min[X_1, \ldots, X_n]$ as the sample minimum.

We then have the following results for m_n and M_n:

- The pdfs and cdfs of M_n and m_n are given by
  - For M_n: $G_n(y) = [F(y)]^n$ and $g_n(y) = nf(y)[F(y)]^{n-1}$
  - For m_n: $H_n(y) = 1 - [1 - F(y)]^n$ and $h_n(y) = nf(y)[1 - F(y)]^{n-1}$
- Thus, we can obtain the moments of M_n and m_n using the above expressions for the pdfs, although in many cases, these calculations are quite difficult.
- Bounds on Expectations: Suppose F has mean and variance μ and $\sigma^2 < \infty$.

  The following are bounds on the means of M_n and m_n:
  - $E\left(M_n\right) \leq \mu + \frac{(n-1)\sigma}{\sqrt{2n-1}}$
  - $E\left(m_n\right) \geq \mu - \frac{(n-1)\sigma}{\sqrt{2n-1}}$
- Asymptotic Results: Let $F_n(y)$ be the cdf of the standardized form of M_n: $\frac{M_n - a_n}{b_n}$,
  - $F_n(y) = P\left[\frac{M_n - a_n}{b_n} \leq y\right]$.
  - $F_n(y) \to G(y)$ as $n \to \infty$ where $G(y)$ is one of the following:
    - $G_{1,\gamma}(y) = e^{-y^{-\gamma}}$ for $y > 0$ and $G_{1,\gamma}(y) = 0$ for $y < 0$
    - $G_{2,\gamma}(y) = e^{-(-y)^{\gamma}}$ for $y < 0$ and $G_{2,\gamma}(y) = 0$ for $y > 0$
    - $G_3(y) = e^{-e^{-y}}$ for $-\infty < y < \infty$
  - If Y has a Weibull distribution, then W = -log(Y) has cdf $G_3(w) = e^{-e^{-(w-\theta_1)/\theta_2}}$ for $-\infty < y < \infty$
  - The sequences (a_n) and (b_n) and the form of $G(y)$ are determined by the form of the cdf $F(y)$ for the data: $Y_1, \ldots, Y_n$. See *A Course in Large Sample Theory*, by Thomas Ferguson for further details.
  - The asymptotic distributions for the sample minimum m_n can be obtained from the results of the sample maximum M_n by using $X_i = -Y_i$.

Sampling Distribution of the Sample Proportion, $\hat{p}$

Suppose we have a population in which all the units in the population are of one of two types: Type A Units or Type B Units. For example, people having a disease or not; or a warehouse containing good parts and defective parts. Let p be the proportion of Type A units in the population. Alternatively, p could be the probability of a particular outcome in a sequence of Bernoulli trials. For example, p is the probability experimental unit no longer has disease after receiving a drug. We apply the drug to n experimental units and observe whether or not they have the disease after a period of time.

Let $\hat{p} = \frac{Y}{n} = \frac{1}{n}\sum_{i=1}^{n} I(Y_i = 1)$ be the sample proportion based on n iid Bernoulli trials or a random sample of n units taken from a population consisting of Type A and Type B units, where $Y_i = 1$ if the ith unit is a Type A unit and $Y_i = 0$ if the ith unit is not a Type A and $Y = \sum_{i=1}^{n} I(Y_i = 1)$ is the total number of Type A outcomes in the sample of n units.

- For sampling with replacement: $n\hat{p}$ has a binomial distribution.
- For sampling without replacement: $n\hat{p}$ has a hypergeometric distribution.
- For both sampling with and without replacement: $E[\hat{p}] = p$.
- For sampling with replacement: $Var[\hat{p}] = \frac{p(1-p)}{n}$
- For sampling without replacement: $Var[\hat{p}] = \left(\frac{N-n}{N-1}\right)\frac{p(1-p)}{n}$, where N is the number of units in the population.

  When $\frac{n}{N} < .05$, $\frac{N-n}{N-1} \approx 1 - \frac{n}{N} > .95 \quad \Rightarrow \quad Var[\hat{p}] \approx \frac{p(1-p)}{n}$
- For large n : The distribution of $\hat{p}$ is approximately $N\left(p, \frac{p(1-p)}{n}\right)$, thus, the asymptotic mean and standard deviation for the sample standard deviation $\hat{p}$ are given by

  $$\mu_A = p \ \text{ and } \ \sigma_A = \frac{\sqrt{p(1-p)}}{\sqrt{n}}$$

  .
- The normal approximation to the binomial distribution should be used only when $min[np, n(1-p)] \geq 5$.

  Also, because we are approximating a discrete distribution with a continuous distribution, the following correction is generally suggested:

  $$Pr[y_1 \leq \hat{p} \leq y_2] \approx \Phi\left(\frac{ny_2 + .5 - np}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{ny_1 - .5 - np}{\sqrt{np(1-p)}}\right),$$

  where Φ is the cdf of a standard normal distribution.

Where does the .5 come from in the approximation?

$$P[\hat{p} = y] = P[n\hat{p} = ny] \approx P[ny - .5 \leq Y \leq ny + .5]$$

where $n\hat{p}$ has a Binomial(n,p) distribution (discrete) and Y has a $N(np, np(1-p))$ distribution which is continuous

Example Let Y have a binomial distribution with n=100, p=0.2

$\mu_Y = E[Y] = np = 100(.2) = 20$ and $\sigma_Y = \sqrt{Var(Y)} = \sqrt{np(1-p)} = \sqrt{100(.2)(.8)} = 4$

Calculate $P(Y \leq 25)$:

- Exact calculation using B(100,.2) distribution:

  $P(Y \leq 25) = pbinom(25, 100, .2) = 0.9125246$

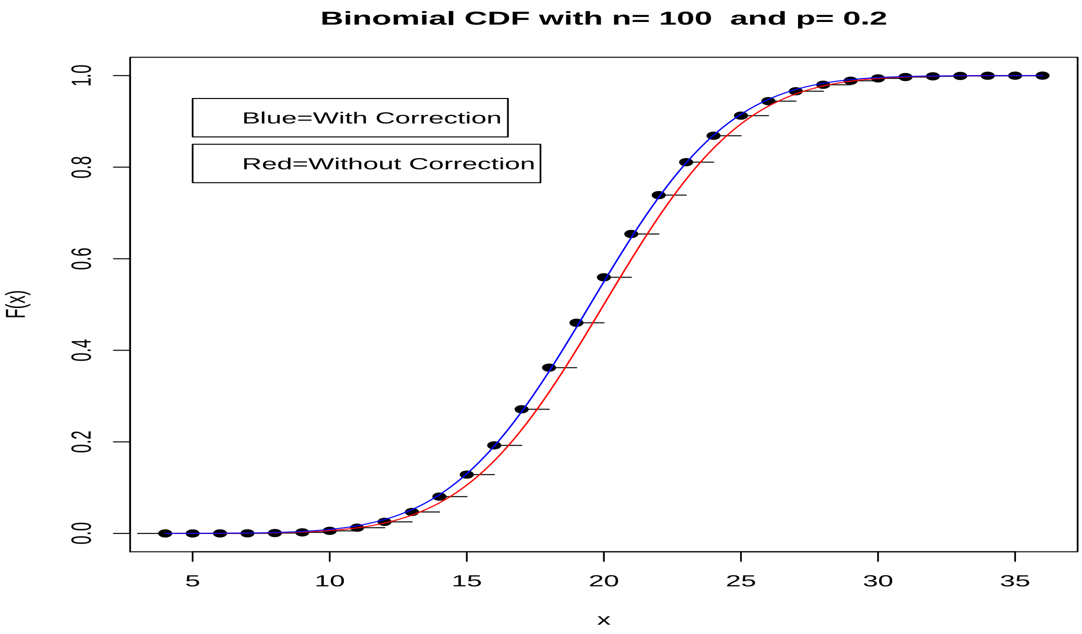
- Normal Approx to Binomial Without Correction:

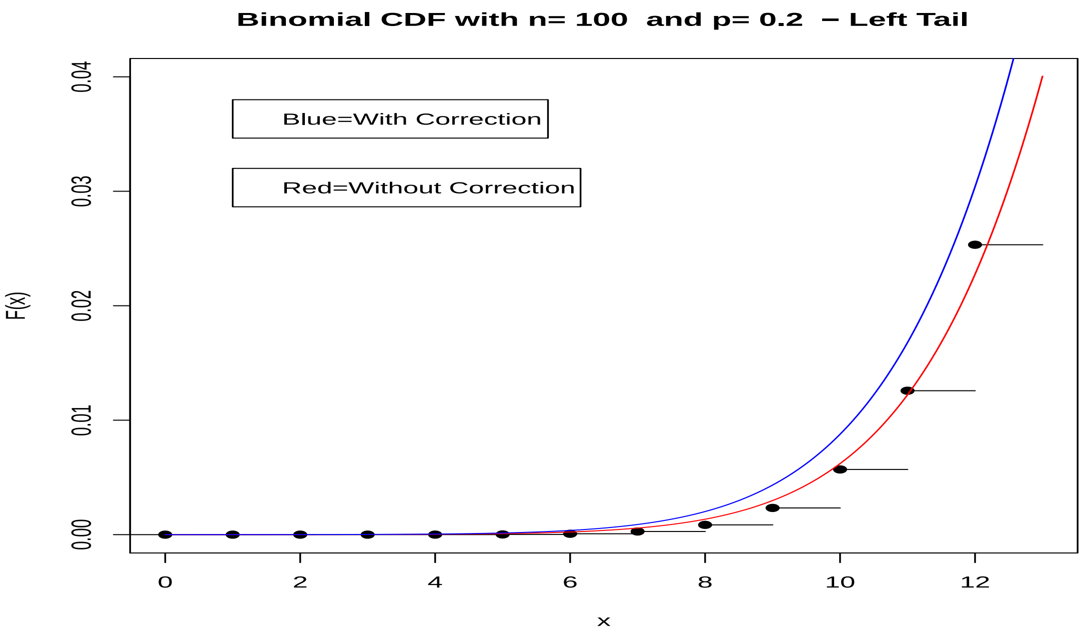
  $P(Y \leq 25) \approx P\left(Z \leq \frac{25-20}{\sqrt{100(.2)(.8)}}\right) = pnorn\left(\frac{25-20}{\sqrt{100(.2)(.8)}}\right) = 0.8943502$

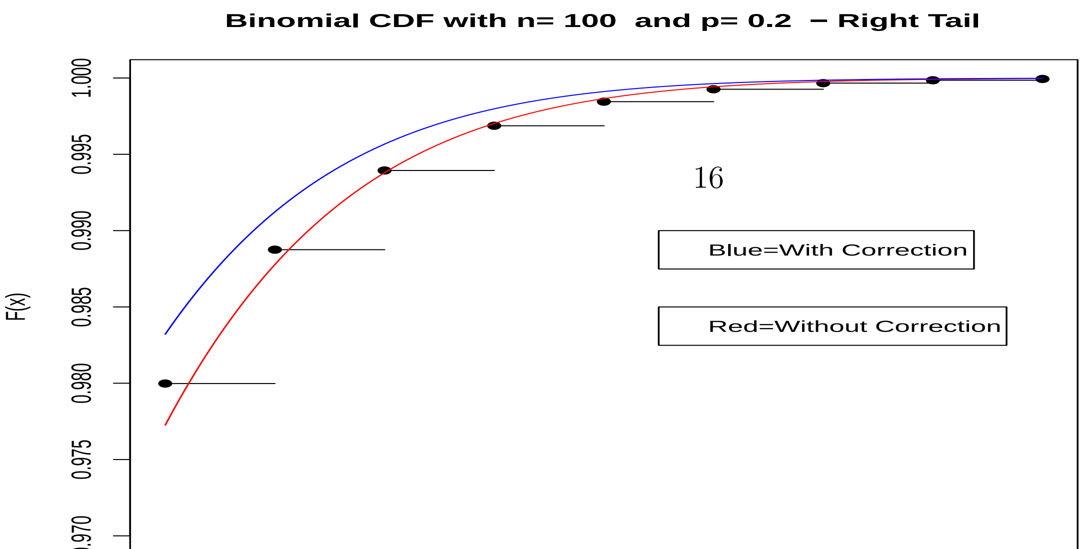
- Normal Approx to Binomial With Correction:

  $P(Y \leq 25) \approx P\left(Z \leq \frac{25-20+.5}{\sqrt{100(.2)(.8)}}\right) = pnorm\left(\frac{25-20+.5}{\sqrt{100(.2)(.8)}}\right) = 0.9154343$

- The following graphs illustrates that for the values of Y in the middle of the binomial distribution the normal approximation with the correction is more accurate than the approximation without the correction. However, in the tails of the distribution, there are regions where the calculation with the correction is less accurate than the calculation without the correction.


Binomial CDF with n= 100 and p= 0.2
Blue=With Correction
Red=Without Correction
F(x)
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
25
30
35
x


Binomial CDF with n= 100 and p= 0.2 – Left Tail
Blue=With Correction
Red=Without Correction
F(x)
0.00
0.01
0.02
0.03
0.04
0
2
4
6
8
10
12
x


Binomial CDF with n= 100 and p= 0.2 – Right Tail
Blue=With Correction
Red=Without Correction
F(x)
0.970
0.975
0.980
0.985
0.990
0.995
1.000

Sampling Distribution of $\bar{X}$, $\hat{Q}(.5)$, S Using Simulation

Using the R program given on the next page, we will simulate data from the standard normal distribution for sample sizes of 10, 25, and 100. For each sample size, 1000 samples of that size are generated. The values of $\bar{X}$ $\hat{Q}(.5)$ S are computed. We thus have for each sample size, 1000 values of $\bar{X}$ $\hat{Q}(.5)$ S. The mean and standard deviation of these 1000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 1000 values will also be provided to demonstrate the shape of the sampling distribution.

Sampling from N(0,1) Distribution

Population parameters:

$\mu = 0, \quad Q(.5) = 0, \quad f(Q(.5)) = f(0) = \frac{1}{\sqrt{2\pi}}, \quad \sigma = 1, \quad \mu_4 = 3\sigma^4 = 3.$

- Asymptotic Mean of $\bar{X}$ is $\mu_A = 0$;
- Asymptotic StDev of $\bar{X}$ is $\sigma_A = \frac{\sigma}{\sqrt{n}} = \frac{1}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $\mu_A = Q(.5) = 0$;
- Asymptotic StDev of $\hat{Q}(.5)$ is $\sigma_A = \frac{\sqrt{(.5)(.5)}}{\sqrt{n}f(Q(.5))} = \frac{(.5)}{\sqrt{n}f(0)} = \frac{(.5)}{\sqrt{n}(1/\sqrt{2\pi})} = \frac{1.2533}{\sqrt{n}}$
- Asymptotic Mean of S is $\mu_A = \sigma = 1$;
- Asymptotic StDev of S is $\sigma_A = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{2}\sigma^2}{2\sigma\sqrt{n}} = \frac{.707}{\sqrt{n}}$

Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0	0.3162	0.0150	0.3132
	25	0	0.2000	-0.0075	0.1975
	100	0	0.1000	0.0058	0.0990
$\hat{Q}(.5)$	10	0	0.3963	0.0185	0.3706
	25	0	0.2507	-0.0043	0.2449
	100	0	0.1253	0.0065	0.1268
S	10	1	0.2236	0.9721	0.2387
	25	1	0.1414	0.9934	0.1417
	100	1	0.0707	1.0003	0.0741

```
#Sampling Distribution of Mean, Median, Standard Deviation
#boxsamdistnorm.R also boxsamdistweib.R and boxsamdisttt5.R are in Dostat/Files/RCode
r=1000
x =  rep(0,10)
m10 =  rep(0,r)
q10 =  rep(0,r)
sq = rep(0,r)
s10 = rep(0,r)
for (i in 1:r){
  x =  rnorm(10)
   m10[i] =  mean(x)
   q10[i] =  median(x)
  s10[i] =  sd(x)}

x =  rep(0,25)
m25 =  rep(0,r)
q25 =  rep(0,r)
s2 = rep(0,r)
s25 = rep(0,r)
for (i in 1:r){
  x=  rnorm(25)
   m25[i] =  mean(x)
   q25[i] =  median(x)
  s25[i] =  sd(x)}

x= rep(0,100)
m100 =  rep(0,r)
q100 =  rep(0,r)
sq = rep(0,r)
s100 = rep(0,r)
for (i in 1:r){
  x=  rnorm(100)
   m100[i] =  mean(x)
   q100[i] =  median(x)
  s100[i] =  sd(x)}

outmean10 = c(mean(m10), mean(q10), mean(s10))
outmean25 = c(mean(m25), mean(q25), mean(s25))
outmean100 = c(mean(m100), mean(q100), mean(s100))
outmean = cbind(outmean10,outmean25,outmean100)
outmean
outsd10 = c(sd(m10), sd(q10), sd(s10))
outsd25 = c(sd(m25), sd(q25), sd(s25))
outsd100 = c(sd(m100), sd(q100), sd(s100))
outsd = cbind(outsd10,outsd25,outsd100)
outsd

postscript("u:/meth1/psfiles/box1sampdistnorm.ps",height=8,horizontal=F)
boxplot(m10,m25,m100,xlab="Sample Size",lab=c(10,10,7),
                  ylab="Sample Mean",
          main="Boxplots of 1000 Sample Means from N(0,1)",
                  names=c("10","25","100"),cex=.75)
postscript("u:/meth1/psfiles/box2sampdistnorm.ps",height=8,horizontal=F)
boxplot(q10,q25,q100,xlab="Sample Size",lab=c(10,10,7),
                  ylab="Sample Median",
          main="Boxplots of 1000 Sample Medians from N(0,1)",
                  names=c("10","25","100"),cex=.75)
postscript("u:/meth1/psfiles/box3sampdistnorm.ps",height=8,horizontal=F)
boxplot(s10,s25,s100,xlab="Sample Size",lab=c(10,10,7),
                  ylab="Sample Standard Deviation",
          main="Boxplots of 1000 of Sample Std Dev from N(0,1)",
                  names=c("10","25","100"),cex=.75)
graphics.off()
```

Sampling from Weibull($\gamma = .61, \alpha = .62$) Distribution

Simulate data from a Weibull distribution with $\alpha = .62, \gamma = .61$ for sample sizes of 10, 25, and 100. This Weibull distribution is highly right skewed. For each sample size, 1000 samples of that size are generated. The values of $\bar{X}$ $\hat{Q}(.5)$ S are computed. We thus have for each sample size, 1000 values of $\bar{X}$ $\hat{Q}(.5)$ S. The mean and standard deviation of these 1000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 1000 values will also be provided to demonstrate the shape of the sampling distribution.

Population parameters: With $\alpha = .62$ and $\gamma = .61$, we obtain

$\mu = \alpha\Gamma\left[(\gamma+1)/\gamma\right] = .9133,\ Q(.5) = \alpha(log(2))^{1/\gamma} = .3400,\ f(Q(.5)) = \alpha(log(2))^{1/\gamma} = .6218,$

$\sigma = \alpha\sqrt{(\Gamma\left[(\gamma+2)/\gamma\right] - (\Gamma\left[(\gamma+1)/\gamma\right])^2)} = 1.5730,\ \mu_4 = 309.2276,\ \ \text{skewness} = 4.38.$

- Mean of $\bar{X}$ is $\mu = .9133$; StDev of $\bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{1.5730}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $Q(.5) = [-(.62)^{.61}log(.5)]^{1/.61} = .3400$;
- Asymptotic StDev of $\hat{Q}(.5) = \frac{\sqrt{(.5)(.5)}}{\sqrt{n}f(Q(.5))} = \frac{(.5)}{\sqrt{n}f(.34)} = \frac{(.5)}{\sqrt{n}(.6218)} = \frac{.8041}{\sqrt{n}}$
- Asymptotic Mean of S is $\sigma = 1.5730$;
- Asymptotic StDev of $S = \frac{\sqrt{\mu_4-\sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{309.2276-(1.573)^4}}{2(1.573)\sqrt{n}} = \frac{5.534}{\sqrt{n}}$

Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0.9133	0.4974	0.9111	0.5010
	25	0.9133	0.3146	0.9106	0.3159
	100	0.9133	0.1573	0.9199	0.1565
$\hat{Q}(.5)$	10	0.3400	0.2543	0.4265	0.3012
	25	0.3400	0.1608	0.3685	0.1657
	100	0.3400	0.0804	0.3491	0.0822
S	10	1.5730	1.7500	1.2795	0.9505
	25	1.5730	1.1068	1.3995	0.7034
	100	1.5730	0.5534	1.5374	0.4411

Sampling from t with df=5 Distribution

Simulate data from a t with df=5 distribution for sample sizes of 10, 25, and 100. This t distribution is symmetric with heavier tables than N(0,1). For each sample size, 1000 samples of that size are generated. The values of $\bar{X}$, $\hat{Q}(.5)$, and S are computed. We thus have for each sample size, 1000 values of $\bar{X}$, $\hat{Q}(.5)$, and S. The mean and standard deviation of these 1000 values are then computed. These means and standard deviations will then be compared to their corresponding asymptotic values based on the central limit theorem. A box plot of the 1000 values will also be provided to demonstrate the shape of the sampling distribution.

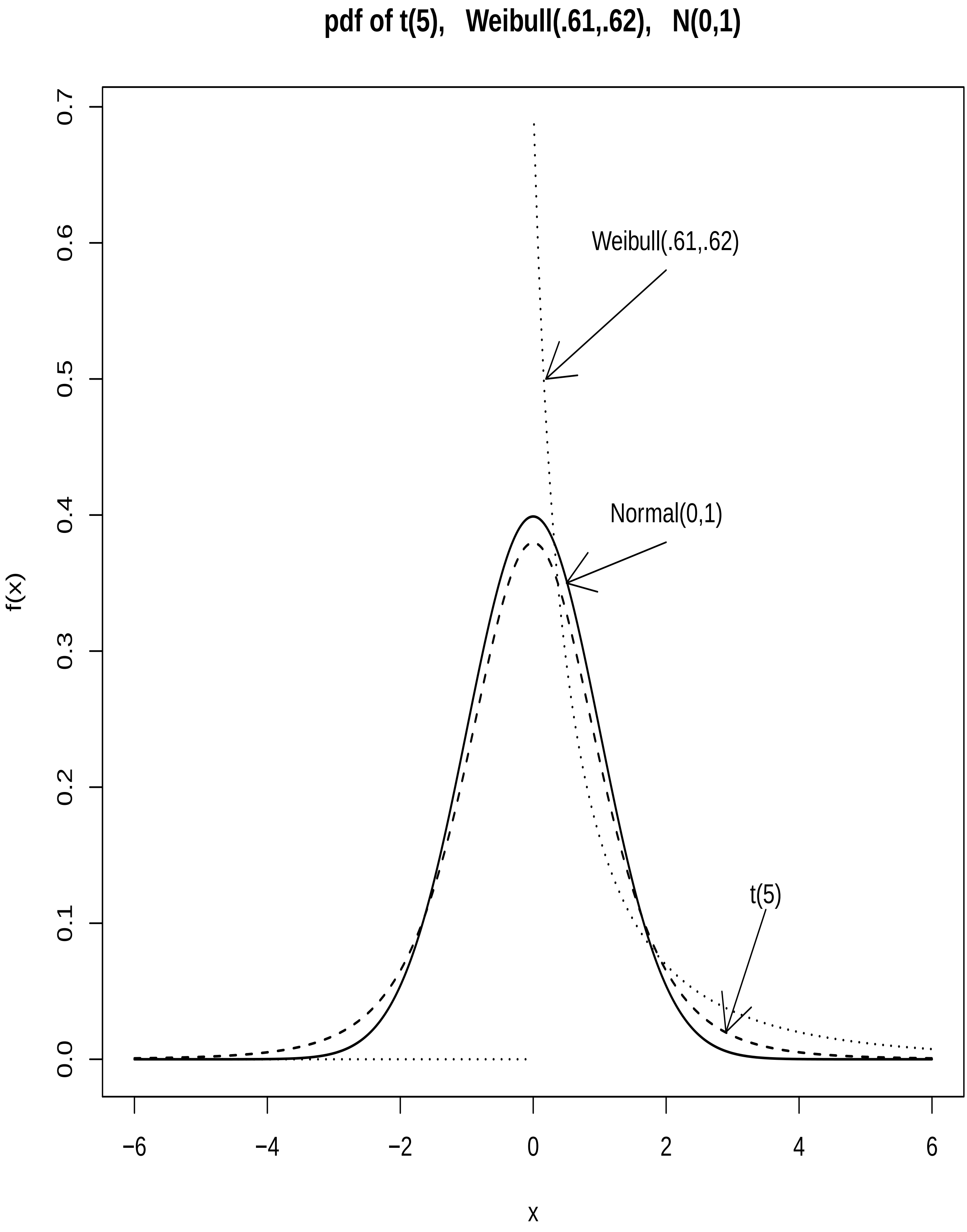
Population parameters: with $\nu = df = 5$

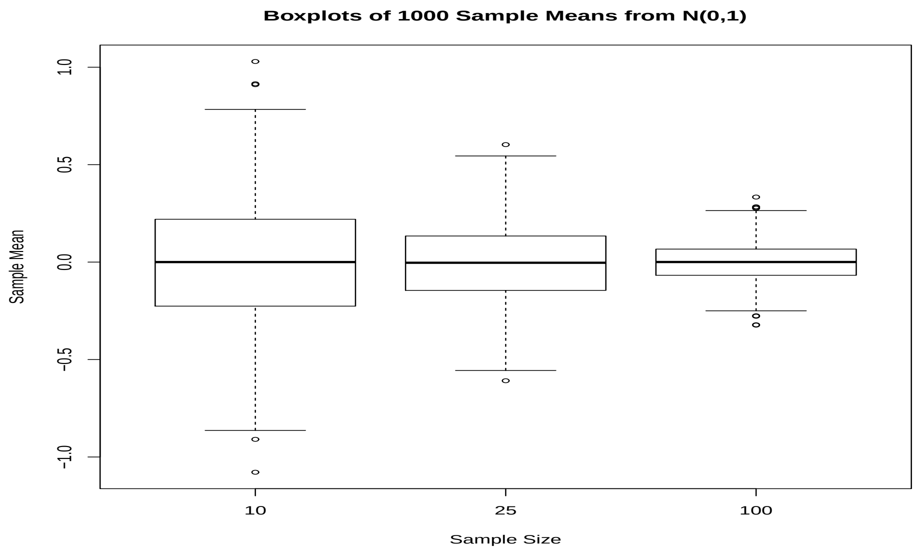
$\mu = 0, \qquad Q(.5) = 0, \qquad f(Q(.5)) = f(0) = \frac{\Gamma((\nu+1)/2)}{\Gamma(\nu/2)} \frac{1}{\sqrt{\nu\pi}} = \frac{8}{3\pi\sqrt{5}} = .3796,$

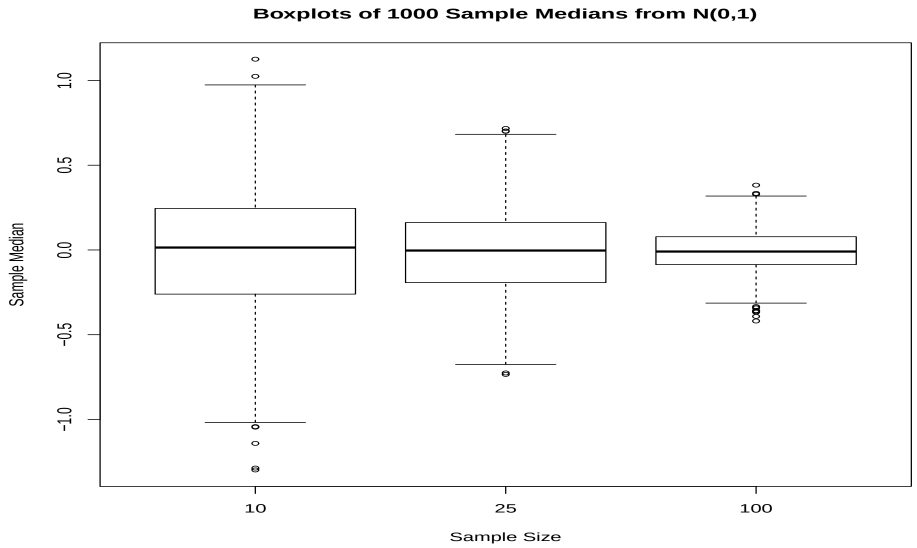
$\sigma = \sqrt{\frac{\nu}{\nu-2}} = \sqrt{\frac{5}{3}} = 1.29099, \qquad \mu_4 = \frac{3\nu^2}{(\nu-2)(\nu-4)} = 25, \qquad \text{skewness} = 0.$

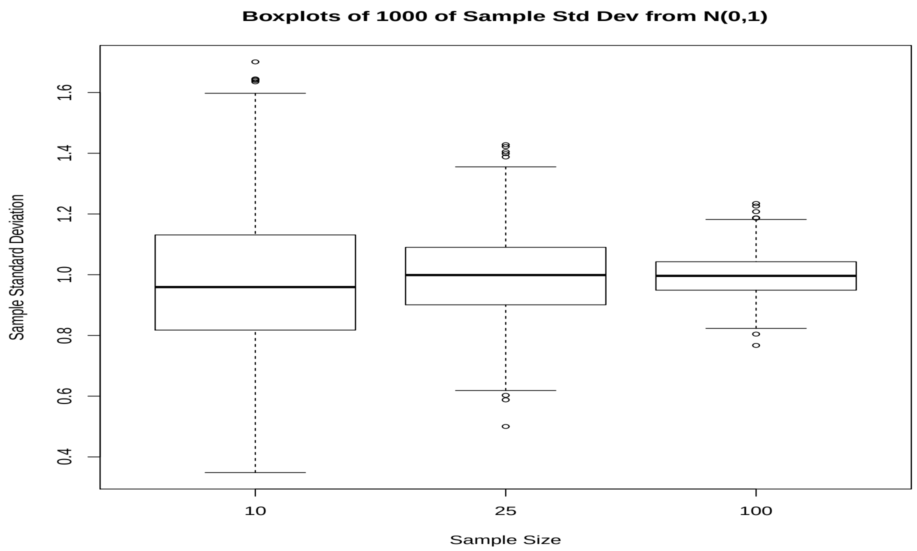
- Mean of $\bar{X}$ is $\mu = 0$;
- StDev of $\bar{X} = \frac{\sigma}{\sqrt{n}} = \frac{1.291}{\sqrt{n}}$
- Asymptotic Mean of $\hat{Q}(.5)$ is $Q(.5) = 0$;
- Asymptotic StDev of $\hat{Q}(.5) = \frac{\sqrt{(.5)(.5)}}{\sqrt{n}f(Q(.5))} = \frac{(.5)}{\sqrt{n}f(0)} = \frac{(.5)}{\sqrt{n}(.3796)} = \frac{1.3172}{\sqrt{n}}$
- Asymptotic Mean of S is $\sigma = 1.291$;
- Asymptotic StDev of $S = \frac{\sqrt{\mu_4 - \sigma^4}}{2\sigma\sqrt{n}} = \frac{\sqrt{25-(1.291)^4}}{2(1.291)\sqrt{n}} = \frac{1.8257}{\sqrt{n}}$

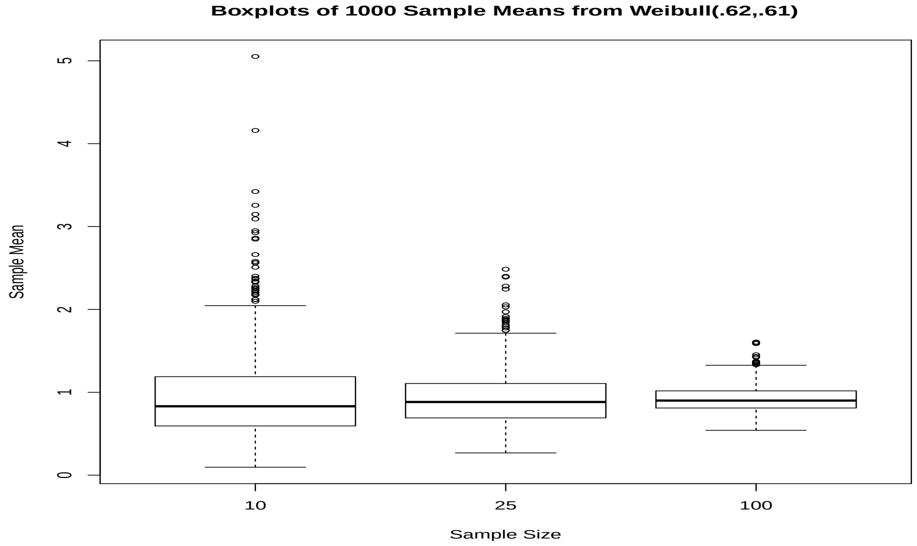
Statistic	Sample Size	Asymp. Mean	Asymp. StDev	Sim. Mean	Sim. StDev
$\bar{X}$	10	0	0.4083	-0.0248	0.3894
	25	0	0.2582	0.0064	0.2567
	100	0	0.1291	0.0024	0.1271
$\hat{Q}(.5)$	10	0	0.4165	-0.0244	0.3914
	25	0	0.2634	-0.0004	0.2672
	100	0	0.1317	0.0019	0.1298
S	10	1.291	0.5773	1.2143	0.3947
	25	1.291	0.3651	1.2676	0.2898
	100	1.291	0.1826	1.2781	0.1558

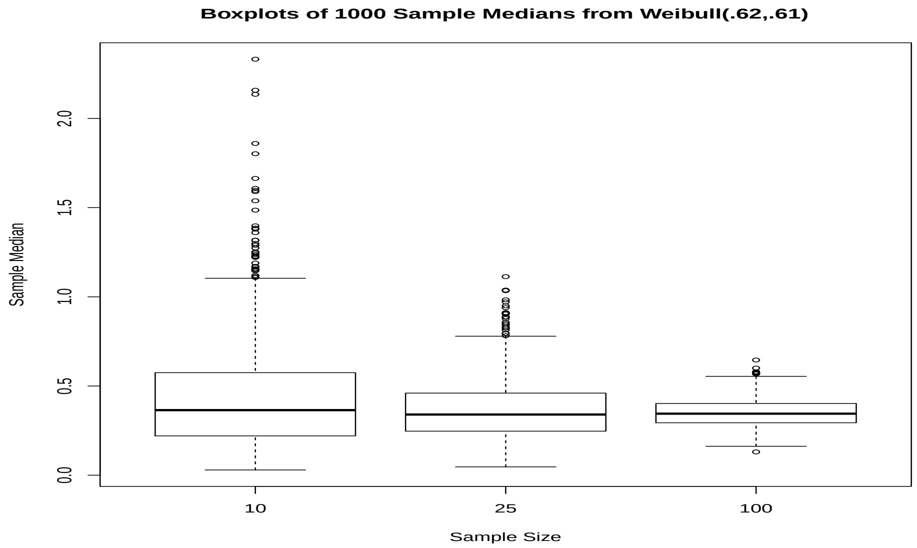

pdf of t(5), Weibull(.61,.62), N(0,1)
Weibull(.61,.62)
Normal(0,1)
t(5)
f(x)
x
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
−6
−4
−2
0
2
4
6

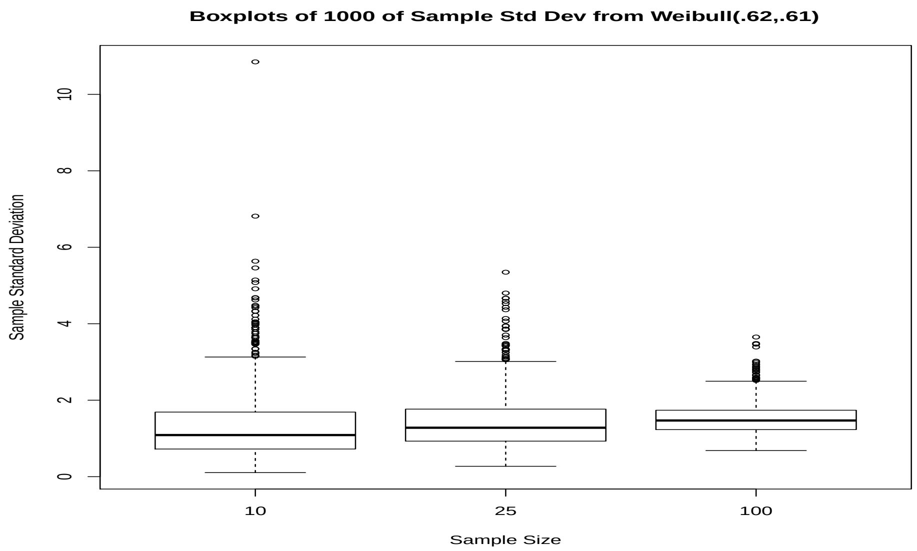

Boxplots of 1000 Sample Means from N(0,1)
Sample Mean
1.0
0.5
0.0
-0.5
-1.0
10
25
100
Sample Size

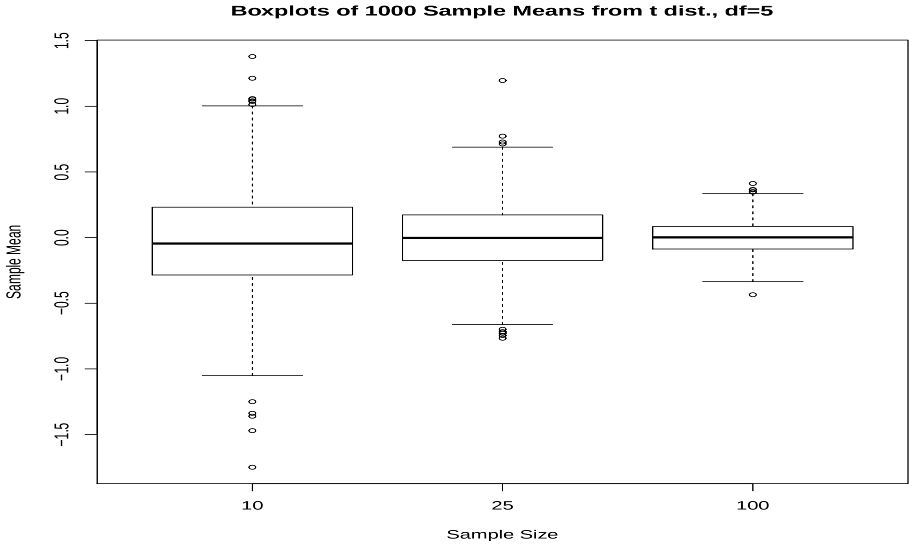

Boxplots of 1000 Sample Medians from N(0,1)
Sample Median
1.0
0.5
0.0
-0.5
-1.0
10
25
100
Sample Size

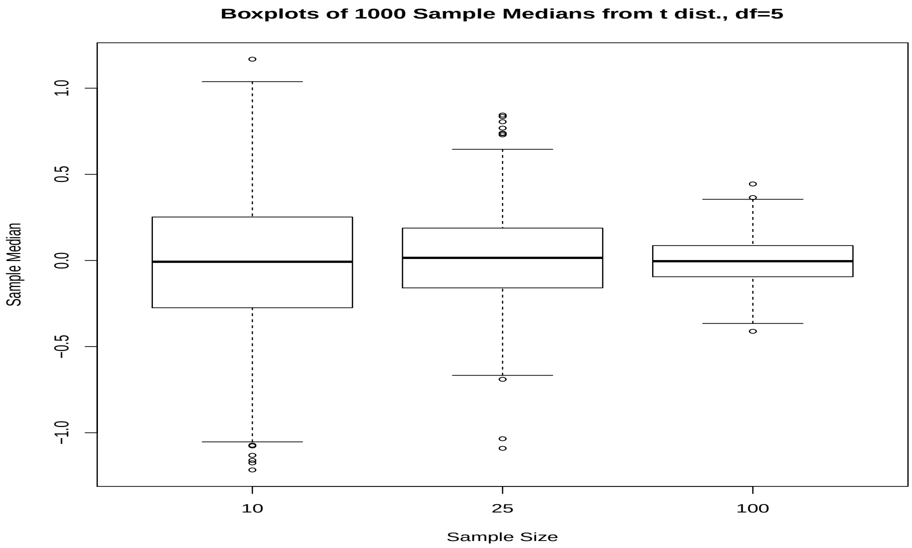

Boxplots of 1000 of Sample Std Dev from N(0,1)
Sample Standard Deviation
1.6
1.4
1.2
1.0
0.8
0.6
0.4
10
25
100
Sample Size

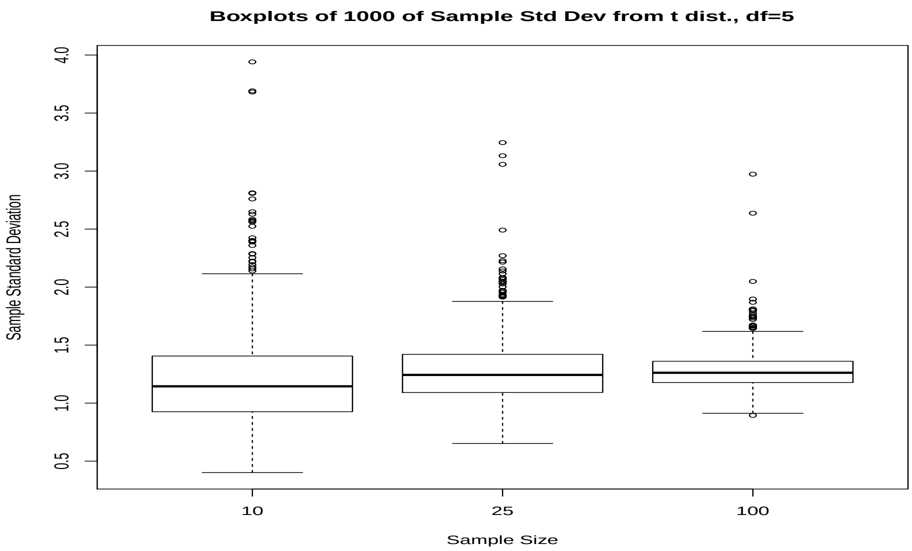

Boxplots of 1000 Sample Means from Weibull(.62,.61)
Sample Mean
0
1
2
3
4
5
10
25
100
Sample Size


Boxplots of 1000 Sample Medians from Weibull(.62,.61)
Sample Median
0.0
0.5
1.0
1.5
2.0
10
25
100
Sample Size


Boxplots of 1000 of Sample Std Dev from Weibull(.62,.61)
Sample Standard Deviation
0
2
4
6
8
10
10
25
100
Sample Size


Boxplots of 1000 Sample Means from t dist., df=5
Sample Mean
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
10
25
100
Sample Size


Boxplots of 1000 Sample Medians from t dist., df=5
Sample Median
1.0
0.5
0.0
−0.5
−1.0
10
25
100
Sample Size


Boxplots of 1000 of Sample Std Dev from t dist., df=5
Sample Standard Deviation
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
10
25
100
Sample Size

Bootstrapping the Sampling Distribution of Various Statistics

Let $X_1, \ldots, X_n$ be iid random variables with a common cdf $F(\cdot)$. Let θ be a parameter which we wish to estimate using a function of the data $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$.

Suppose the cdf $F(\cdot)$ is unknown and the sample size n is small or that the asymptotic distribution of $\hat{\theta}$ is intractable.

We wish to determine the sampling distribution of $\hat{\theta}$ in order to be able to determine its bias as an estimator of θ or its variance or its percentiles in order to provide an assessment of how well $\hat{\theta}$ estimates θ. Suppose we can not mathematically derive the true sampling distribution of $\hat{\theta}$ because the cdf $F(\cdot)$ is unknown or the form of $\hat{\theta}$ may be too complex to obtain an exact result. The asymptotic distribution may not provide an adequate approximation of the true sampling distribution of $\hat{\theta}$ because n is too small. If F were known we could use simulation to determine the sampling distribution of $\hat{\theta}$.

For example, suppose we wanted to determine the sampling distribution of an estimate of the scale parameter, θ, in the Cauchy distribution when the location parameter is 0 and the sample size is $n = 100$. We would first select various values for θ. Next, for each value of θ, simulate $N = 10,000$ sets of 100 realizations from the Cauchy distribution, and compute a value for $\hat{\theta}$: $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_N$,. We could then use these N values of $\hat{\theta}$ to estimate the sampling distribution of $\hat{\theta}$ for the selected value of θ. This would then be repeated for the various choices of θ. However, in many cases, F will be unknown.

An alternative to these approaches which can be used when F is unknown is the bootstrap procedure which will provide an approximation to the sampling distribution of $\hat{\theta}$ in the situation where we can write θ as a function of the cdf, that is, $\theta = g(F(\cdot))$.

For example,

- the population mean $\mu = \int_{-\infty}^{\infty} x dF(x)$,
- the population variance, $\sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 dF(x)$,
- the population median M can be defined by $.5 = \int_{-\infty}^{M} dF(x)$.

To obtain the sample estimator, we simply replace the true cdf $F(\cdot)$ with the empirical (sample) cdf $\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x) = \frac{\#\{X_i \leq x\}}{n}$
in $\theta = g(F(\cdot))$ to obtain $\hat{\theta} = g(\hat{F}(\cdot))$.

In order to obtain the sampling distribution of $\hat{\theta}$, we simulate data from the edf $\hat{F}(\cdot)$ in place of the true cdf $F(\cdot)$. Recall, we used the true cdf in the simulations from the normal and Weibull distributions. That is, we will now consider the population to be the observed data having a cdf which places mass $\frac{1}{n}$ on each of the observed data values X_i. Thus, we select M random samples of size n (sampling with replacement) from this "new" population, compute $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$. We now have M realizations of $\hat{\theta}$ from which we can estimate the pdf (using a kernel density estimator), the quantile function, or specific parameters like its mean: $\mu_{\hat{\theta}} = E[\hat{\theta}] = \frac{1}{M}\sum_{i=1}^{M} \hat{\theta}_i$ and its variance $Var[\hat{\theta}] = \frac{1}{M}\sum_{i=1}^{M}(\hat{\theta}_i - \mu_{\hat{\theta}})^2$. Similarly, we can compute its median or any other percentiles.

When using a bootstrap procedure, we have two levels of approximation:

1. The estimation of the population cdf F using the edf $\hat{F}$

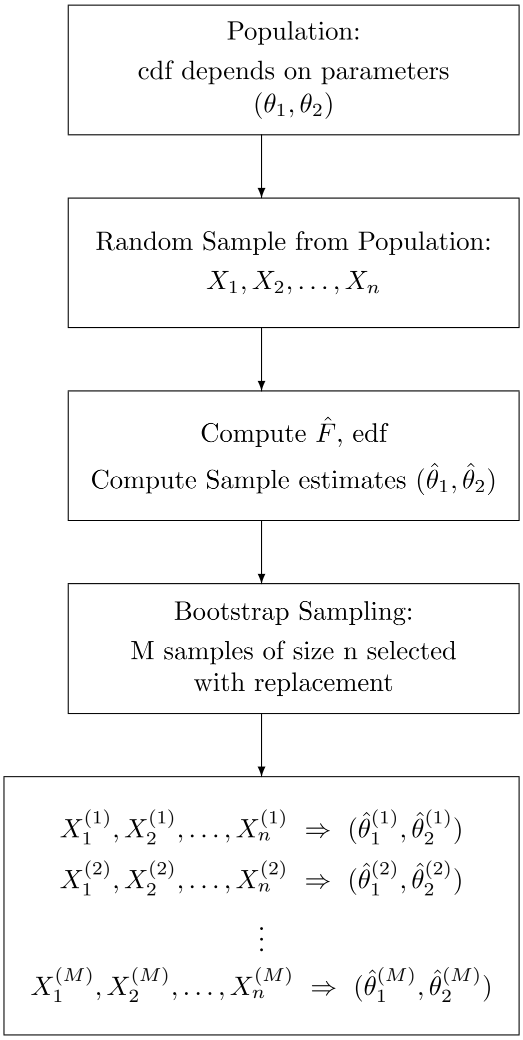
   Accuracy of approximation controlled by n

2. Repeated estimation of the edf $\hat{F}$ using the M bootstrap estimators $\tilde{F}$

   Accuracy of approximation limited by the value of n and how well $\hat{F}$ approximates F.

A flow chart of bootstraped procedure is given here:

1. Obtain data $X_1, \ldots, X_n$ iid with cdf $F(\cdot)$
2. Compute $\hat{\theta} = \hat{\theta}(X_1, \ldots, X_n)$
3. Select a random sample of size n with replacement from $X_1, \ldots, X_n$ (i.e., simulate n independent observations from the edf $\hat{F}(\cdot)$): Denote by $X_1^*, \ldots, X_n^*$
4. Compute $\hat{\theta}_i^* = \hat{\theta}(X_1^*, \ldots, X_n^*)$
5. Repeat Step 3 and Step 4 M times yielding $\hat{\theta}_1^*, \hat{\theta}_2^*, \ldots, \hat{\theta}_M^*$
6. Use these M realizations of $\hat{\theta}$ to construct the sampling distribution of $\hat{\theta}$: Means, StDev, Percentiles, pdf, etc.



We will consider the following example to illustrate the application of the bootstrap procedure:

EXAMPLE: Suppose the life lengths of eleven engine parts are measured as

5700, 36300, 12400, 28000, 19300, 21500, 12900, 4100, 91400, 7600, 1600

We want to estimate the median life length θ of the engine part. From the data we compute $\hat{\theta} = X_{(6)} = 12900$. To study the variation in this estimator we need to know its sampling distribution. We will use the bootstrap to approximate this distribution.

We will first generate M = 200 bootstrap samples from $\hat{F}(\cdot)$ and then M = 20,000 bootstrap samples using the following R code: **bootsampMed.R**

```
y = c(1600,4100,5700,7600,12400,12900,19300,21500,28000,36300,91400)
mhat = median(y)
M =  20000
d = numeric(M)
for(i in 1:M)d[i] = median(sample(y,replace=T))
hist(d)
bootmean =  mean(d)
bootstd = sd(d)
bootquant = quantile(d)
postscript("bootexample20000",height=8,horizontal=F)
probs = seq(0,1,.01)
Qd = quantile(d,probs)
boxplot(d,main="Empirical Quantile for Sample Median",
ylab="Median Life Lengths of Engine Part",plot=T)
plot(probs,Qd,type="l",ylab="Q(u) for Median",xlab="u",xlim=c(0,1),lab=c(10,11,7))
title("Empirical Quantile for Sample Median",cex=.75)
plot(density(d),type="l",xlab="Median Life Lengths",ylab="PDF of Sample Median")
title("Empirical pdf for Sample Median",cex=.75)
qqnorm(d,main="Normal Prob Plot of Sample Median",
xlab="normal quantiles",ylab="Sample Medians",
lab=c(7,7,7),cex=.75)
qqline(d)
graphics.off()
```

In the following table the first five simulations are given with a "+" indicating which of the original data values was sampled. Note that some values will be sampled multiple times and some values may not be included:

First Five Bootstrap Samples

Original Data	Bootstrap Sample				
	1	2	3	4	5
1600			+		++
4100	+++	++	+	+	
5700	+	+	+	+++	+
7600				++	++
12400	+	+	+		+
12900	+		++		
19300	+	+	+	++	
21500		+++	+	+	+
28000	+	++	+	++	+
36300	+	+			+
91400	++		++		++
$\hat{\theta}^*$	12900	21500	12900	7600	12400

From the 200 realizations of $\hat{\theta}^*$, the following summary statistics were computed:

Average: $E[\hat{\theta}^*] = \frac{1}{200}\sum_{i=1}^{200} \hat{\theta}_i^* = 14877.5$

Standard Deviation: $\sqrt{Var[\hat{\theta}^*]} = \sqrt{\frac{1}{200}\sum_{i=1}^{200}(\hat{\theta}_i^* - E[\hat{\theta}^*])^2} = 5552.6$

Quantile				
0	.25	.50	.75	1.0
4100	12400	12900	19300	36300

If we extended the simulation to 20000 bootstrap samples, we obtain

Average: $$E[\hat{\theta}^*] = \frac{1}{20000} \sum_{i=1}^{20000} \hat{\theta}_i^* = 14924.1$$

Standard Deviation: $$\sqrt{Var[\hat{\theta}^*]} = \sqrt{\frac{1}{20000} \sum_{i=1}^{20000} (\hat{\theta}_i^* - E[\hat{\theta}^*])^2} = 5933.7$$

Quantile				
0	.25	.50	.75	1.0
1600	12400	12900	19300	91400

Thus, there was only minor differences in the mean and standard deviation for the sampling distribution of the median when comparing 200 bootstrap samples to 20000 bootstrap samples. However, note the big discrepancies between the quantiles. When generating 20000 samples of size 11 from the original data set, samples were obtained in which the median of the bootstrap sample was equal to the minimum value (1600) in the original data set. Because the bootstrap median equals $\hat{\theta}^* = X^*_{(6)}$, this result implies that, in the bootstrap samples having median=1600, at least 6 of the 11 data values must be equal to 1600. This seems very unlikely. However, if we calculate the expected number of samples in the 20000 samples having exactly 6 of their 11 values equal to 1600, we find that

$$\text{Expected Number} = 20000\ P[\text{exactly 6 of 11 values equal 1600}]$$

$$= (20000) \left(\binom{11}{6} (1)^6 (10)^5 \right) / (11)^{11} = 3.2$$

Therefore, on the average we would expect 3.2 occurrences of the event that exactly 6 of the 11 data values were equal to 1600.

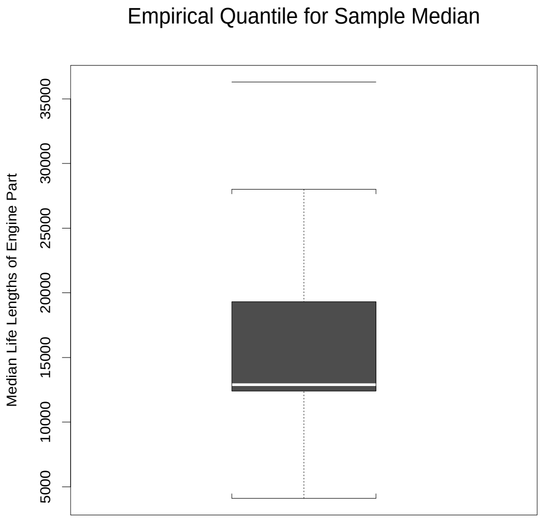
A good reference on bootstraping is *Bootstrap Methods and their Applications* by D.V. Hinkley and A.C. Davison.

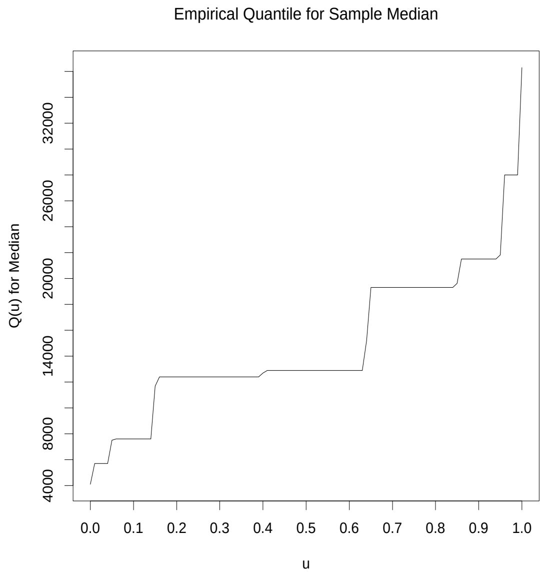
A plot of the quantile function, kernel density estimator of the pdf, a box plot, and a normal reference distribution plot for the sampling distribution of the sample quantile are given on the next pages for 200 and 20000 bootstrap samples. We note that there is considerable differences in the plots. The plots for 20000 bootstrap samples reveals the discreteness of the possible values for the median when the sample size ($n = 11$ in our case) is very small. Also, we note that $n = 11$ is too small for the sampling distribution for the median to achieve its asymptotic result (n large), an approximate normal distribution.

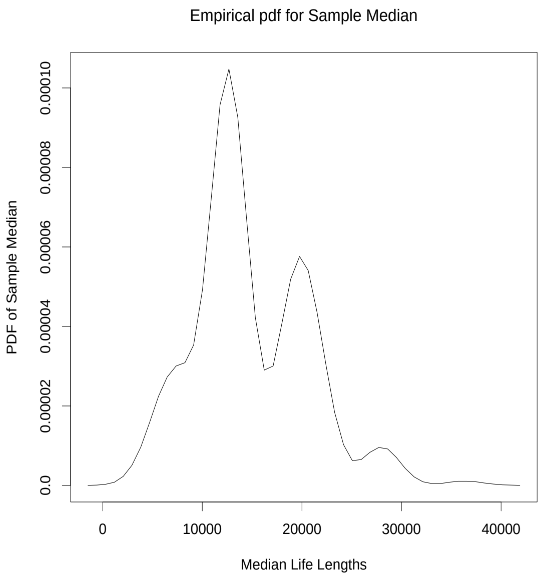
If you wanted to observe a few of the bootstrap samples, say the first K=20, then just use the following R code:

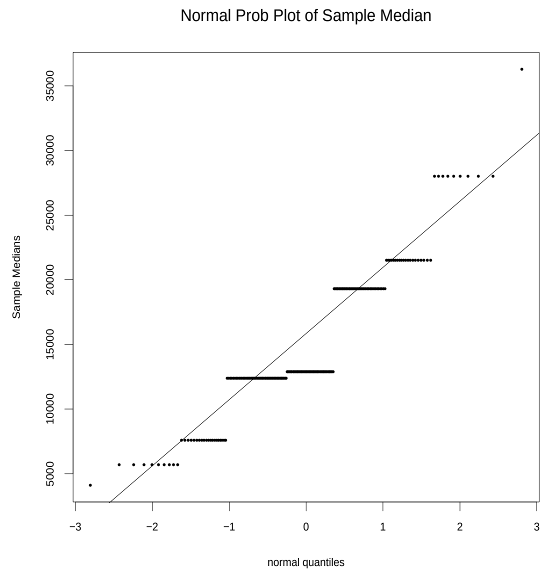
```
y = c(1600,4100,5700,7600,12400,12900,19300,21500,28000,36300,91400)
n = length(y)
K = 20
sam = matrix(0,K,n)
for(i in 1:K)
{
sam[i,] = sample(y,replace=T)
}
```

Plots of sampling distribution of sample median based on 200 resamples.

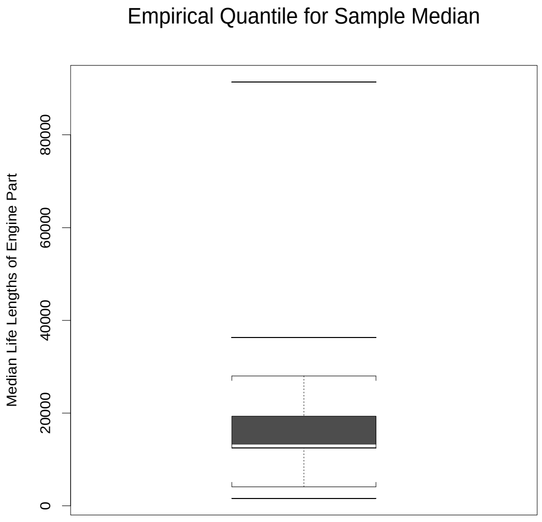


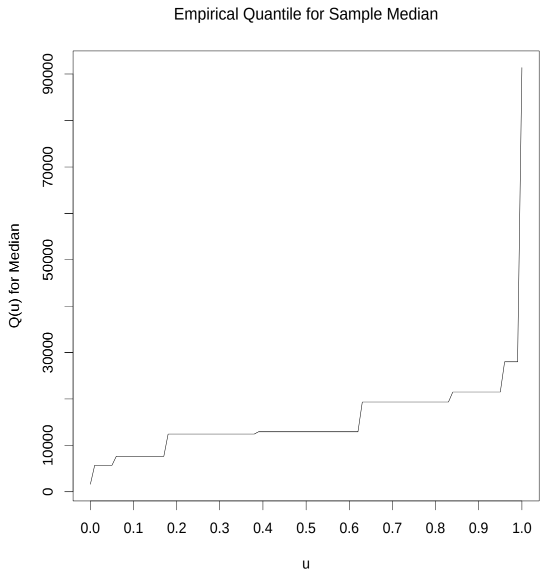


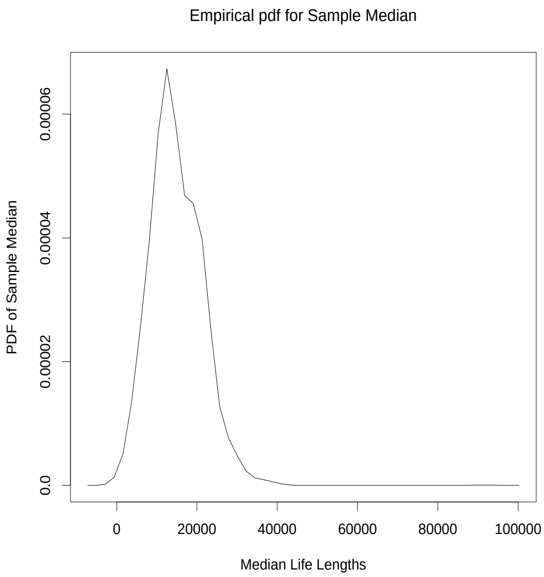


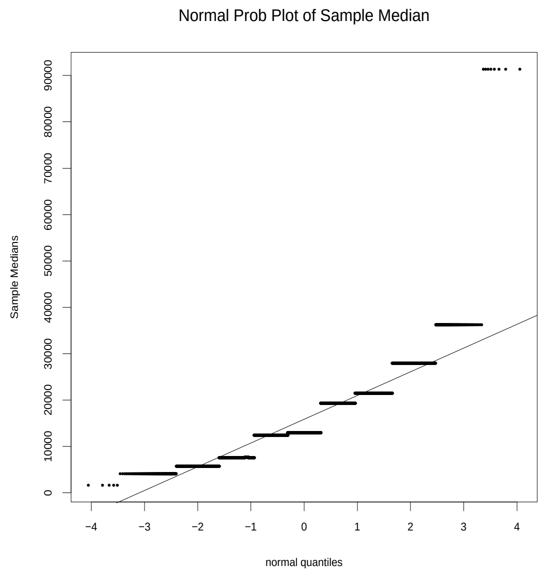


Plots of sampling distribution of sample median based on 20000 resamples.









Sampling Distribution of Maximum Likelihood Estimator

Let $Y_1, Y_2, \ldots, Y_n$ be a random sample (or iid observations) from a population/process having pdf $f(y)$ which depends on unknown parameters: θ_1, θ_2.

- The MLEs of the θs is that vector $(\widehat{\theta}_1, \widehat{\theta}_2)$ which maximizes the likelihood function:

$$L(\widehat{\theta}_1, \widehat{\theta}_2) = \max_{\theta \epsilon \Theta} L(\theta_1, \theta_2)$$

- Invariance Property of MLE:

  If $\hat{\theta}$ is the MLE of θ, then for any function $h(\theta)$, the MLE of $h(\theta)$ is $h(\hat{\theta})$

  Example: Suppose random sample is from a Weibull population and $\hat{\beta}$ and $\hat{\gamma}$ are the MLEs. Find the MLE of $S(t) = e^{-t^{\gamma}/\beta}$

  Solution: By the Invariance Property of MLEs, the MLE of $S(t)$ is

  $\hat{S}(t) = e^{-t^{\hat{\gamma}}/\hat{\beta}}$

- The fixed n properties of MLEs depend on the population pdf $f(\cdot)$ and no general statement can be made about the distribution of $\widehat{\theta}$

- Asymptotic Properties of MLEs:

  Under some regularity conditions (see Casella-Berger,2nd Edition, p516), the central limit theorem for MLEs yields

  Let $\hat{\theta}$ denote the MLE of θ. Let $h(\theta)$ be any continuous function of θ,

  For large n : The distribution of $h(\hat{\theta})$ is approximately $N\left(h(\theta), \frac{(h'(\theta))^2}{I_n(\theta)}\right)$, where $I_n(\theta) = E_\theta\left(-\frac{\partial^2}{\partial\theta^2} log(L(\theta))\right)$.

  Thus, the asymptotic mean and standard deviation for the sampling distribution of $h(\hat{\theta})$ are given by

$$\mu_A = h(\theta) \qquad\qquad \sigma_A = \frac{h'(\theta)}{\sqrt{I_n(\theta)}},$$

an estimator of σ_A is given by

$$\widehat{\sigma}_A = \frac{h'(\hat{\theta})}{\sqrt{I_n(\hat{\theta})}}$$

In particular, if $h(\theta) = \theta$, then the MLE of θ, $\hat{\theta}$ is approximately (large n) normally distributed with asymptotic mean and standard deviation

$$\mu_A = \theta \qquad\qquad \sigma_A = \frac{1}{\sqrt{I_n(\theta)}}$$

- Example 1: Let $f(\cdot)$ be exponential(β). We derived in Handout 6 that the MLE was $\hat{\beta} = \bar{Y}$

  To find the asymptotic variance, we need to find the form of the Information number:

$$L(\beta) = \prod_{i=1}^{n} \frac{1}{\beta} e^{-t_i/\beta} = \beta^{-n} e^{-\frac{1}{\beta}\sum_{i=1}^{n} t_i}$$

$$l(\beta; y) = log[L(\beta)] = -nlog(\beta) - \frac{1}{\beta}\sum_{i=1}^{n} t_i$$

$$\frac{\partial l(\beta; y)}{\partial \beta^2} = -\frac{n}{\beta} + \frac{1}{\beta^2}\sum_{i=1}^{n} t_i \qquad\qquad \frac{\partial^2 l(\beta; y)}{\partial \beta^2} = \frac{n}{\beta^2} - \frac{2}{\beta^3}\sum_{i=1}^{n} t_i$$

$$I_n(\beta) = E\left[-\frac{\partial^2 l(\beta; y)}{\partial \beta^2}\right] = -\frac{n}{\beta^2} + \frac{2}{\beta^3} E\left[\sum_{i=1}^{n} t_i\right] \quad = -\frac{n}{\beta^2} + \frac{2n\beta}{\beta^3} \quad = \frac{n}{\beta^2}$$

  Thus, the asymptotic standard deviation is approximated by

$$\widehat{\sigma}_A = \frac{1}{\sqrt{I_n(\widehat{\beta})}} = \frac{\widehat{\beta}}{\sqrt{n}}$$

- Example 2: Let $f(\cdot)$ be Weibull(γ,α). We demonstrated in Handout 6 that the MLE estimators are obtained from solving the equations:

$$\widehat{\alpha} = \left(\frac{1}{n}\sum_{i=1}^{n} t_i^{\widehat{\gamma}}\right)^{1/\hat{\gamma}}$$

$$\frac{\sum_{i=1}^{n} t_i^{\widehat{\gamma}} log(t_i)}{\sum_{i=1}^{n} t_i^{\widehat{\gamma}}} - \frac{1}{\widehat{\gamma}} = \frac{1}{n}\sum_{i=1}^{n} log(t_i)$$

  To find the asymptotic variances we would need to approximate numerically the second partial derivatives of the log-likelihood function at the values of $\hat{\gamma}$ and $\hat{\alpha}$.

  These are the values displayed in the R output that were provided in Handout 6.

```
y = c(15.321, 9.008, 20.104, 7.729, 45.154, 8.404, 5.332, 0.577, 4.305, 4.517,
12.594, 6.829, 3.291, 37.175, 0.841, 1.317, 7.613, 20.582, 2.030, 10.001,
4.666, 12.933, 0.591, 39.454,  8.875)

library(MASS)
fitdistr(y,"weibull")


OUTPUT from R:

     shape         scale
  0.9839245   11.4852981
( 0.1512936) ( 2.4660607)
```

  We thus have

  $\hat{\gamma} = 0.9739245$ with estimated standard error: $\widehat{SE}(\hat{\gamma}) = 0.1512936$

  $\hat{\alpha} = 11.4852981$ with estimated standard error: $\widehat{SE}(\hat{\alpha}) = 2.4660607$

Parametric Bootstrap

To obtain the small sample sampling distribution of the MLE in those situations where the form of the MLE makes the exact derivation intractable or n is too small to invoke the asymptotic results, we can use parametric bootstrap techniques.

Let f be the pdf of the population or process that generates the data. Suppose f depends on unknown parameters $\theta_1, \theta_2, \ldots, \theta_k$. To simplify the notation, let $k = 1$.

Let $Y_1, Y_2, \ldots, Y_n$ be iid with pdf $f(y, \theta)$.

Obtain the MLE of θ, $\hat{\theta}$

To obtain the sampling distribution of $\hat{\theta}$, make use of parametric bootstrap:

1. Generate M samples of size n from $f(y, \hat{\theta})$ and from each sample obtain a value for $\hat{\theta}$

   $$y_{11}, y_{12}, \ldots, y_{1n} \quad \Rightarrow \quad \hat{\theta}_1$$

   $$y_{12}, y_{22}, \ldots, y_{2n} \quad \Rightarrow \quad \hat{\theta}_2$$

   $$y_{13}, y_{32}, \ldots, y_{3n} \quad \Rightarrow \quad \hat{\theta}_3$$

   $$\vdots$$

   $$y_{1M}, y_{M2}, \ldots, y_{Mn} \quad \Rightarrow \quad \hat{\theta}_M$$

2. Use the values of $\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_M$ to estimate percentiles and moments of the sampling distribution of $\hat{\theta}$, for example,

$$\widehat{Q}(.05), \qquad \widehat{Q}(.95), \qquad \widehat{Q}(.5)$$

$$\hat{\mu}_{\hat{\theta}} = \frac{1}{M} \sum_{i=1}^{M} \hat{\theta}_i, \qquad \hat{\sigma}_{\hat{\theta}} = \sqrt{\frac{1}{M-1} \sum_{i=1}^{M} (\hat{\theta}_i - \hat{\mu}_{\hat{\theta}})^2}$$

Example

A study is conducted to evaluate the time to failure of a new device. From previous studies, the reliability engineer is fairly certain that the time to failure, T, will have a Weibull distribution with parameters γ, and α, but the parameters are unknown because this a new device. The researcher wants to estimate $S(t) = e^{-(t/\alpha)^\gamma}$ for $t = 20$ and hence needs to know the sampling distribution of the MLE of $S(20) = P[T > 20]$, the proportion of the devices produced that will fail after 20 units of time.

$$\hat{S}(20) = e^{-(t/\hat{\alpha})^{\hat{\gamma}}} \text{ where } \hat{\alpha} \text{ and } \hat{\gamma} \text{ are the MLE's of } \alpha \text{ and } \gamma$$

The researcher was only able to test 25 of the devices and obtain their times to failure: $T_1, T_2, \ldots, T_{25}$ but $n = 25$ is too small to invoke the asymptotic distribution of $\hat{S}(20)_{MLE}$.

Therefore, a parametric bootstrap approximation is a possible alternative:

1. Use R to obtain values for the MLE's, $(\hat{\gamma}_{MLE}, \hat{\alpha}_{MLE})$ based on the 25 data values.
2. Use the computed values of ($\hat{\gamma}_{MLE}$ and $\hat{\alpha}_{MLE}$) to generate 1000 samples of size $n = 20$ from a Weibull distribution and then compute an estimate of $S(20)$ using $\hat{S}(20)$ from each of the 1000 samples yielding

$$\hat{S}(20)_1, \ \hat{S}(20)_2, \ \ldots, \ \hat{S}(20)_{1000}$$

- In R, iterate the function $rweibull(20, \hat{\gamma}_{MLE}, \hat{\alpha}_{MLE})$ $M = 1000$ times to obtain

  Sample 1: $y_{1,1}, y_{1,2}, \ldots, y_{1,25} \ \Rightarrow \ (\hat{\gamma}_1, \hat{\alpha}_1) \ \Rightarrow \ \hat{S}(20)_1 = e^{-(20/\hat{\alpha}_1)^{\hat{\gamma}_1}}$

  Sample 2: $y_{2,1}, y_{2,2}, \ldots, y_{2,25} \ \Rightarrow \ (\hat{\gamma}_2, \hat{\alpha}_2) \ \Rightarrow \ \hat{S}(20)_2 = e^{-(20/\hat{\alpha}_2)^{\hat{\gamma}_2}}$

  $\vdots$

  Sample 1000: $y_{1000,1}, y_{1000,2}, \ldots, y_{1000,25} \ \Rightarrow \ (\hat{\gamma}_{1000}, \hat{\alpha}_{1000}) \ \Rightarrow \ \hat{S}(20)_{1000} = e^{-(20/\hat{\alpha}_{1000})^{\hat{\gamma}_{1000}}}$

- Use the 1000 values of $\hat{S}(20)$: $\hat{S}(20)_1, \hat{S}(20)_2, \ldots, \hat{S}(20)_{1000}$ to estimate the standard error and necessary percentiles of $\hat{S}(20)$

The following are the times to failure of the 25 devices:

```
0.577  0.591  0.841  1.317  2.030  3.291  4.305  4.517  4.666  5.332
6.829  7.613  7.729  8.404  8.875  9.008 10.001 12.594 12.933 15.321
20.104 20.582 37.175 39.454 45.154
```

If we did not have a model for the data, a distribution-free estimate would be $\hat{S}(20) = P[T > 20] = 5/25 = .2$

The following R code can be used to obtain the parametric bootstrap estimates:

```
x =  c(15.321, 9.008, 20.104, 7.729, 45.154, 8.404, 5.332, 0.577, 4.305, 4.517,
  12.594, 6.829, 3.291, 37.175, 0.841, 1.317, 7.613, 20.582, 2.030, 10.001,
  4.666, 12.933, 0.591, 39.454,  8.875)
n =  length(x)
library(MASS)
fitdistr(x,"weibull")
#  OUTPUT from R:
#
#      shape        scale
#    0.9839245   11.4852981
#  ( 0.1512936) ( 2.4660607)
gamma = 0.9839245
alpha = 11.4852981
B  =  1000
W  =  matrix(0,B,n)
A =  numeric(B)
A =  rep(0,B)
G =  numeric(B)
G =  rep(0,B)
S = numeric(B)
S = rep(0,B)
{
for (i in 1:B)
W[i,]  =  rweibull(n,gamma,alpha)
}
{
for (i in 1:B)
G[i] = fitdistr(W[i,],"weibull")$estimate[1]
}
{
for (i in 1:B)
A[i] = fitdistr(W[i,],"weibull")$estimate[2]
}
{
for (i in 1:B)
S[i] = exp(-(20/A[i])^G[i])
}
summary(S)
sd(S)
boxplot(S)
out=c(mean(G),sd(G),mean(A),sd(A))
```

Running the R-code yields the following results:

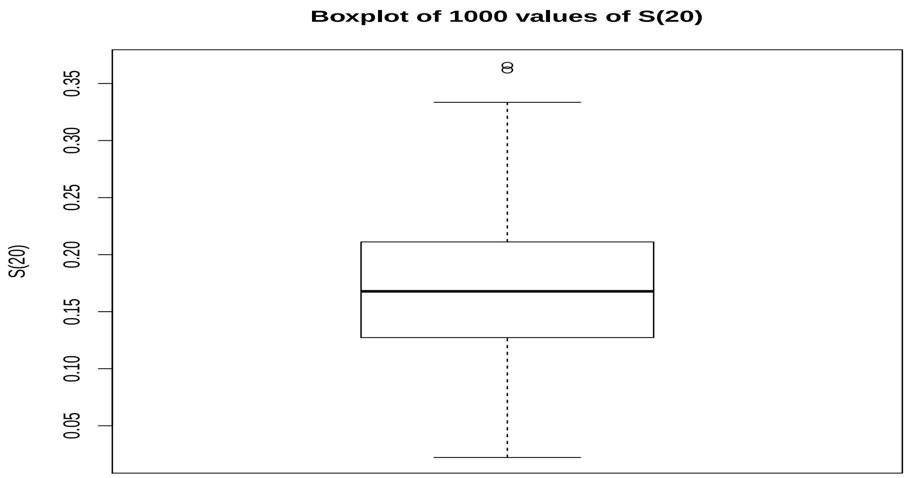
```
summary(S)

  Min.      1st Qu.    Median     Mean      3rd Qu.     Max.
0.008895   0.133400   0.173200   0.174700  0.213800   0.418300

sd(S)
 0.06236332

boxplot(S)
out=c(mean(G),sd(G),mean(A),sd(A))
out

  1.0502193  0.1844123 11.6973586  2.4982325
```



The average value of the 1000 values of $\hat{S}(20)$ was 0.1747 which is nearly identical to our point estimator from the data using the MLE's:

$$\hat{\alpha} = 11.4852981, \quad \hat{\gamma} = .9839245 \quad \Rightarrow$$

$$\hat{S}(20) = e^{-(20/11.4852981)^{.9839245}} = .178013$$

The standard error of the parameters using the asymptotic formulas are

$\widehat{se}(\hat{\gamma}) = .1512936$ and $\widehat{se}(\hat{\alpha}) = 2.4660607$

which are very close to our bootstrap estimates 0.1844123 and 2.4982325, even though n=25 is not very large.

Finally, we have the estimated standard error of $\hat{S}(20)$ is 0.06236332 which can be used to place a C.I. on S(20). It would be a fairly complicated computation to obtain the asymptotic standard error from MLE theory.