START T. 10/22/27

HANDOUT #11: INTERVAL ESTIMATORS

I. Confidence Intervals (CI) for a Parameter θ

(a) Pivot Method
(b) Exact CI
(c) Asymptotic CI
(d) Bootstrap CI
(e) Example of Improper Use of CLThm
(f) Sample Size Determination
(g) Distribution-Free CI for Q(u)

II. Prediction Intervals (PI)

(a) PI for Normal Population Distribution
(b) PI for Exponential Population Distribution

III. Tolerance Intervals (TI)

(a) TI for Normal Population Distribution
(b) Lower Tolerance Bound for Exponential Population Distribution
(c) Distribution-Free Tolerance Bounds

IV. Alternative Approaches to CI, PI, TI

(a) Transformations
(b) Bootstraping

Supplemental Reading

- Sections 6.1, 6.2, 7.1-7.4, 9.1, 9.3, 14.6, and pages 565-566 in Tamhane/Dunlop book

Interval Estimators

Suppose we have a population or process having unknown parameters θ. For example, the population mean μ, standard deviation σ, population proportion p, or population median $Q(.5)$. Alternatively, we may know that the population cdf $F(\cdot;\theta)$ is a member of a family of distributions but θ is unknown. For example, Weibull(θ,γ) or Poisson(λ). We would like to estimate various components of the population. Three such interval estimators are

1. $100(1-\alpha)$ **Confidence Interval on the parameter θ:** Plausible set of values for θ

   a. Based on a random sample $Y_1,\ldots,Y_n$ from a population or process, obtain a point estimator $\hat{\boldsymbol{\theta}}$: MLE, MOM, Robust

   b. Construct interval of values $(\hat{\theta}_L,\hat{\theta}_U)$ such that $P[\hat{\theta}_L \le \theta \le \hat{\theta}_U] = 1-\alpha$.

   c. The $100(1-\alpha)$ C.I. $(\hat{\theta}_L,\hat{\theta}_U)$ reflects the uncertainty in using $\hat{\theta}$ as an estimator of θ.

2. $100(1-\alpha)$ **Prediction Interval on the R.V. Y:**

   a. Let $Y_1,\ldots,Y_n$ be a random sample from a population or process.

   b. Based on the data, predict the value of the next r.v. Y_{n+1} selected from the population or process: $\hat{Y}_{n+1}$.

   c. The $100(1-\alpha)$ P.I. is an interval of values $(\hat{Y}_{n+1,L},\hat{Y}_{n+1,U})$ such that $P[\hat{Y}_{n+1,L} \le Y_{n+1} \le \hat{Y}_{n+1,U}] = 1-\alpha$

3. (P,γ) **Tolerance Interval on the Population or Process:**

   Based on a random sample $Y_1,\ldots,Y_n$ from a population or process, construct an interval of values $(L_{p,\gamma},U_{p,\gamma})$ such that we are $100\gamma\%$ certain that the interval $(L_{p,\gamma},U_{p,\gamma})$ contains at least $100P\%$ of the population values.

   In $L_{p,\gamma}$, p is the proportion of the population values to be contained in the interval and γ is the level of confidence that the interval will in fact contain $100p\%$ of the population values.

Distinct differences in the three types of intervals:

1. The C.I. is making an inference about a fixed population parameter, μ, σ, or a parameter in a family of distributions, for example, β in an exponential family or λ in a Poisson family. For example, we are 99% certain that the mean time to appearance of a tumor is in the interval (40, 60) hours.

2. The P.I. is forecasting or predicting the value of a R.V., for example, we are 98% certain that the tensile strength of the next specimen of alloy tested is in the interval (1.23, 2.31) units. A meteorologist predicts with 80% confidence that the amount of rainfall tomorrow will be .5 to 1 cm.

3. The Tolerance Interval is estimating a region in a population that contains at least $100P\%$ of the population values. For example,

    a. Suppose based on the data we compute an interval of values such that we are 95% certain that 99% of 1 cm bearings produced next month by Company X will have diameters in the region $(.99, 1.02)$ cm. This is a $(P = .99, \gamma = .95)$ Tolerance Interval on distribution for the diameter of ball bearings.

    b. Suppose we compute from the data a lower bound such that we are 80% certain that at least 95% of all new Honda SUV's produced next year will have miles to failure of its transmission of at least 120000 miles. Then, $(120000, \infty)$ would be a $(P = .95, \gamma = .80)$ Lower Tolerance Bound on the miles to failure distribution.

Illustration of Level of Confidence In many samples 95% of intervals would contain μ

- 100 random samples of size $n = 25$ were generated from a population having a $N(27, (.8)^2)$ distribution:

$$Y_{i,1}, Y_{i,2}, \ldots, Y_{i,25}, \quad i = 1, 2, \ldots, 100$$

- From each of the 100 samples, a 95% C.I. for μ was constructed:

$$C.I._i = \bar{Y}_i \pm 1.96\frac{.8}{\sqrt{25}} \quad i = 1, 2, \ldots, 100$$

- There are now 100 estimates of μ: $C.I._{.1}, C.I._{.2}, \ldots, C.I._{.100}$.

- How many of the 100 intervals contain μ?

  The graph on the next page provides the answer to this question.

  This illustrates the relative frequency interpretation of a confidence interval.

- In repeated sampling (many more than 100), the proportion of $100(1-\alpha)\%$ confidence intervals that contain the population parameter is $(1-\alpha)$.

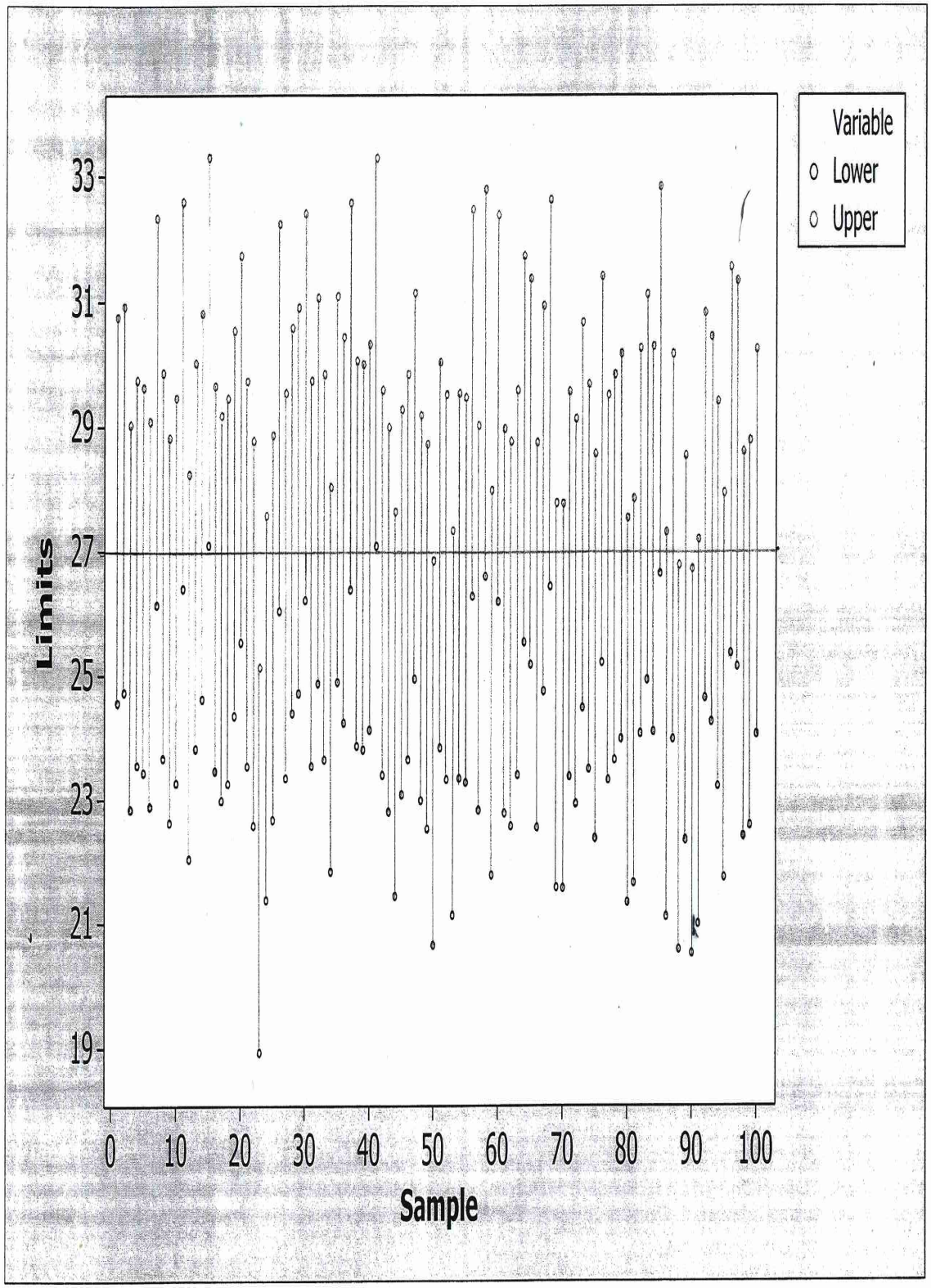
- The proportion of the intervals which **fail** to contain the population parameter is α.

  In the example on the next page, we have $6/100 = .06 \approx .05$ intervals fail to contain $\mu = 27$.

- However, the probability that a realized 95% C.I., (25.31, 25.95), for the population mean, actually contains μ is either 0 or 1.

  The value $100(1-\alpha)\%$ is relative to repeated sampling or relative to the proportion over a very large number of such intervals not to a particular interval.

- The level of confidence $100(1-\alpha)\%$ refers to the process of constructing the confidence interval and not to the actual confidence interval constructed from a given data set. That is, the process used to obtain the 95% C.I. (25.31, 25.95) for the population mean generates intervals of which 95% of the C.I.s produced will contain the population mean and 5% will not contain the population mean.


Variable
Lower
Upper
Limits
33
31
29
27
25
23
21
19
0
10
20
30
40
50
60
70
80
90
100
Sample

Construction of C.I.'s

The Pivot Method will often be used to construct C.I.'s for population or distribution parameters.

1. Find a function of the data $Y = (Y_1, \ldots, Y_n)$ and the parameter θ, $g(Y, \theta)$, such that the sampling distribution of $g(Y, \theta)$ does not depend on θ.

2. Use the cdf of $g(Y, \theta)$, $H(\cdot)$, to determine two percentiles $C_{\frac{\alpha}{2}}$ and $C_{1-\frac{\alpha}{2}}$ satisfying
   $P[C_{\frac{\alpha}{2}} \leq g(Y, \theta) \leq C_{1-\frac{\alpha}{2}}] = 1 - \alpha$, that is,
   $H(C_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$ and $H(C_{1-\frac{\alpha}{2}}) = 1 - \frac{\alpha}{2}$,
   where H is the cdf of the pivot.

3. Invert the inequality $C_{\frac{\alpha}{2}} \leq g(Y, \theta) \leq C_{1-\frac{\alpha}{2}}$ to obtain

   $L(C, g(Y, \theta)) \leq \theta \leq U(C, g(Y, \theta))$.

   Thus conclude that $(L(C, g(Y, \theta)), U(C, g(Y, \theta)))$ is a $100(1 - \alpha)\%$ C.I. for θ.

Determining the Distribution of Pivot

After we select the appropriate pivot, the cdf $H(\cdot)$ of the sampling distribution of the pivot must be determined in order to obtain $C_{\frac{\alpha}{2}}$ and $C_{1-\frac{\alpha}{2}}$. The methods for determining these distributions involve

1. Exact mathematical derivation of the distribution of the pivot

2. Using the asymptotic distribution of the pivot - Wald C.I.

3. Using a bootstrap approximation to the distribution of the pivot

1. Mathematical Determination of Distribution of Pivot

EXAMPLE 1. C.I. for μ for $N(\mu, \sigma^2)$ Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$

Pivot is $g(Y, \mu) = \frac{\bar{Y}-\mu}{S/\sqrt{n}}$

which is a function of the data $\bar{Y}$, S and the unknown parameter μ.

The sampling distribution of $g(Y, \mu)$ is a $t-$distribution with $df = n-1$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu, \sigma)$.

Next we obtain the percentiles

$C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1-\frac{\alpha}{2}, n-1)$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1-\frac{\alpha}{2}, n-1)$, where

$t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n-1$. That is,

$$1-\alpha = P[-t_{\frac{\alpha}{2}} \leq \frac{\bar{Y}-\mu}{S/\sqrt{n}} \leq t_{\frac{\alpha}{2}}] \Rightarrow \quad \bar{Y} - t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}} \Rightarrow$$

$$\bar{Y} \pm t_{\frac{\alpha}{2}}\frac{S}{\sqrt{n}}$$

is a $100(1-\alpha)\%$ C.I. for μ when population distribution is $N(\mu, \sigma^2)$.

EXAMPLE 2. C.I. for σ for $N(\mu, \sigma^2)$ Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$

Pivot is $g(Y, \sigma) = \frac{(n-1)S^2}{\sigma^2}$

which is a function of the data through S and the unknown parameter σ.

The sampling distribution of $g(Y, \sigma)$ is a chi-square distribution with $df = n-1$ which does not depend on the unknown parameter, $\boldsymbol{\theta} = \sigma$.

Next we obtain the percentiles $C_{\frac{\alpha}{2}} = \chi^2_{\frac{\alpha}{2}} = qchisq(\frac{\alpha}{2}, n-1)$ and

$C_{1-\frac{\alpha}{2}} = \chi^2_{1-\frac{\alpha}{2}} = qchisq(1-\frac{\alpha}{2}, n-1)$,

which are the lower and the upper $\frac{\alpha}{2}$ percentile of the chi-square distribution with $df = n-1$. That is,

$$1-\alpha = P\left[\chi^2_{\frac{\alpha}{2}} \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\frac{\alpha}{2}}\right] \Rightarrow \quad \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}} \Rightarrow$$

$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2}}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2}}}}\right)$$

is a $100(1-\alpha)\%$ C.I. for σ when population distribution is $N(\mu, \sigma^2)$.

EXAMPLE 3. C.I. for $\mu_2 - \mu_1$ for comparing two $N(\mu_i, \sigma_i^2)$ Distribution with $i = 1,2$

Let $X = (X_1, \ldots, X_{n_1})$ be iid $N(\mu_1, \sigma_1^2)$ and $Y = (Y_1, \ldots, Y_{n_2})$ be iid $N(\mu_2, \sigma_2^2)$, with X's and Y's independent.

Case 1: $\sigma_1 = \sigma_2 = \sigma$ (with σ unknown)

Pivot is $g(Y, X, \mu_1, \mu_2) = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}$,

where $S_p^2 = \frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_2+n_1-2}$ is the weighted (pooled) estimator of σ^2.

The sampling distribution of $g(Y, X, \mu_1, \mu_2)$ is a $t-$distribution with $df = n_1 + n_2 - 2$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu_1, \mu_2, \sigma_1, \sigma_2)$.

Next we obtain the percentiles $C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1 - \frac{\alpha}{2}, n_1 + n_2 - 2)$,

where $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n_1 + n_2 - 2$. That is,

$$1 - \alpha = P\left[-t_{\frac{\alpha}{2}} \leq \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_p^2}{n_1}+\frac{S_p^2}{n_2}}} \leq t_{\frac{\alpha}{2}}\right] \Rightarrow$$

$$(\bar{Y} - \bar{X}) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \leq \mu_2 - \mu_1 \leq (\bar{Y} - \bar{X}) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}} \Rightarrow$$

$$\bar{Y} - \bar{X} \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

is a $100(1 - \alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $Y_i's$ and $X_i's$ are independent r.v's from population distributions $N(\mu_i, \sigma^2), i = 1, 2$.

Case 2: $\sigma_1 \neq \sigma_2$

Let $X = (X_1, \ldots, X_{n_1})$ be iid $N(\mu_1, \sigma_1^2)$ and $Y = (Y_1, \ldots, Y_{n_2})$ be iid $N(\mu_2, \sigma_2^2)$, with X's and Y's independent.

with $\sigma_1 \neq \sigma_2$, hence do not use pooled estimator

Pivot is $g(Y, X, \mu_1, \mu_2) = \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}}$,

where S_i^2 are the sample variances and estimators of σ_i^2.

The sampling distribution of $g(Y, X, \mu_1, \mu_2)$ is not exactly a t-distribution but Welch-Satterthwaite proved that the distribution is approximately a $t-$distribution with

$df_{WS} = \frac{(w_1+w_2)^2}{\frac{w_1^2}{n_1-1}+\frac{w_2^2}{n_2-1}}$,

where $w_i = \frac{S_i^2}{n_i}, i = 1, 2.$

Next we obtain the percentiles

$C_{\frac{\alpha}{2}} = -t_{\frac{\alpha}{2}} = -qt(1-\frac{\alpha}{2}, df_{WS})$ and $C_{1-\frac{\alpha}{2}} = t_{\frac{\alpha}{2}} = qt(1-\frac{\alpha}{2}, df_{WS})$,

where $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = df_{WS}$.
That is,

$$1-\alpha \approx P\left[-t_{\frac{\alpha}{2}} \leq \frac{(\bar{Y}-\bar{X})-(\mu_2-\mu_1)}{\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}} \leq t_{\frac{\alpha}{2}}\right] \quad \Rightarrow$$

$$(\bar{Y}-\bar{X}) - t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \leq \mu_2-\mu_1 \leq (\bar{Y}-\bar{X}) + t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \quad \Rightarrow$$

$$\bar{Y}-\bar{X} \pm t_{\frac{\alpha}{2}}\sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}}$$

is an approximate $100(1-\alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $Y_i's$ and $X_i's$ are independent r.v's from population distributions $N(\mu_i, \sigma_i^2), i = 1, 2.$

Case 3: Y_i and X_i paired

Let $(Y_1, X_1), \ldots, (Y_n, X_n)$ be iid pairs of observations.

The goal is to construct a C.I. on $\mu_Y - \mu_X$.

The problem with using the pivot for the case were the X's and Y's were independent is displayed as follows:

When using a paired analysis, the pairing reduces the variability of the individual differences if the $(X, Y)-$pairs are positively correlated:

$$Var(\bar{Y} - \bar{X}) = Var(\bar{Y}) + Var(\bar{X}) - 2\sigma_X \sigma_Y Corr(\bar{Y}, \bar{X}) < Var(\bar{Y}) + Var(\bar{X})$$

provided $Corr(\bar{Y}, \bar{X}) > 0$

In many experiments involving paired observations, the sample size n is relatively small and hence an estimate of $Corr(\bar{Y}, \bar{X})$ is not feasible. Thus, the data is reduced to the differences in the n pairs:

Let $D_i = Y_i - X_i,\ i = 1, \ldots, n$ with $D_i's$ iid $N(\mu_D, \sigma_D^2)$, and $\mu_D = \mu_2 - \mu_1$.

Pivot is

$$g(D, \mu_D) = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

which is a function of the data $\bar{D}$, S_D and the unknown parameter μ_D.

The sampling distribution of $g(Y, \mu)$ is a $t-$distribution with $df = n - 1$ which does not depend on the unknown parameters $\boldsymbol{\theta} = (\mu_D, \sigma_D)$. We have converted the problem to finding a C.I. for the mean of a normal population. Therefore,

$$\bar{D} \pm t_{\frac{\alpha}{2}} \frac{S_D}{\sqrt{n}}$$

is a $100(1 - \alpha)\%$ C.I. for $\mu_2 - \mu_1$ when $D_i's$ are normally distributed.

Note: $t_{\frac{\alpha}{2}}$ is the upper $\frac{\alpha}{2}$ percentile of the t-distribution with $df = n - 1$, not $df = 2(n - 1)$.

When using a paired analysis, the pairing reduces the variability of the individual differences if the $(X, Y)-$pairs are positively correlated:

$$Var(\bar{D}) = Var(\bar{Y} - \bar{X}) = Var(\bar{Y}) + Var(\bar{X}) - 2Cov(\bar{Y}, \bar{X}) < Var(\bar{Y}) + Var(\bar{X})$$

The reduction of the variance would lead to a C.I. of narrow width, and hence a more precise estimation of $\mu_2 - \mu_1$, if a paired data experiment was conducted rather than having two independent data sets. However, the reduction in variance is obtained only by a concurrent reduction in the df for the $t-$percentiles. That is,

$$t_{\frac{\alpha}{2}, n-1} > t_{\frac{\alpha}{2}, 2(n-1)}$$

Thus, in attempting to decide between using a paired analysis or two independent samples, this trade-off between reduced variability and reduced df must be taken into consideration.

EXAMPLE 4. C.I. for β in an Exponential Distribution

Let $Y_1, \ldots, Y_n$ be iid Exp(β)

The MLE of β is $\hat{\beta} = \bar{Y}$ which will be used to form the Pivot:

Pivot is $g(Y, \beta) = \frac{2n\bar{Y}}{\beta}$.

Verify that the sampling distribution of $g(Y, \beta)$ does not depend on β.

Recall, $n\bar{Y} = \sum_{i=1}^{n} Y_i$ has a $Gamma(\alpha = n, \beta)$ distribution.

Further recall, that if W has a $Gamma(\alpha = n, \beta)$ distribution then $X = 2W/\beta$ has a chi-square distribution with $df = 2n$.

Therefore, the sampling distribution of the pivot $g(Y, \beta) = \frac{2n\bar{Y}}{\beta}$ is chi-square with df=2n.

It then follows that

$1 - \alpha = P\left[\chi^2_{\frac{\alpha}{2}} \leq \frac{2n\bar{Y}}{\beta} \leq \chi^2_{1-\frac{\alpha}{2}}\right]$,

where $\chi^2_{\frac{\alpha}{2}} = qchisq(\frac{\alpha}{2}, 2n)$ and $\chi^2_{1-\frac{\alpha}{2}} = qchisq(1 - \frac{\alpha}{2}, 2n)$

are respectively the lower and upper $\frac{\alpha}{2}$-percentiles of a chi-square distribution with $df = 2n$.

Inverting the inequalities yields a $100(1 - \alpha)\%$ C.I. for β :

$$\left(\frac{2n\bar{Y}}{\chi^2_{1-\frac{\alpha}{2}}}, \frac{2n\bar{Y}}{\chi^2_{\frac{\alpha}{2}}}\right)$$

The tables in most textbooks provide the upper percentiles of the chi-square distribution.

For example, if we wanted a 95% C.I. with n=10, then we would select the following values from the table:

$\chi^2_{\frac{\alpha}{2}} = \chi^2_{.025,20} = 9.591$ using $\alpha = 1 - .025 = .975$ in Chisquared table or use **qchisq(.025,20** in R

$\chi^2_{1-\frac{\alpha}{2}} = \chi^2_{1-.025,20} = 34.170$ using $\alpha = .025$ in Chisquared table or use **qchisq(.975,20** in R

EXAMPLE 5. C.I. for $\theta = \beta_1/\beta_2$ for comparing Two Exponential cdf's

Let $Y = (Y_1, \ldots, Y_{n_2})$ be iid Exp(β_2) and $X = (X_1, \ldots, X_{n_1})$ be iid Exp(β_1) with $Y_i's$ and $X_i's$ independent.

Pivot is $g(Y, X, \beta_1, \beta_2) = \frac{\bar{Y}/\beta_2}{\bar{X}/\beta_1} = \frac{\left(2n_2\bar{Y}/\beta_2\right)/(2n_2)}{\left(2n_1\bar{X}/\beta_1\right)/(2n_1)}$

The sampling distribution of $g(Y, X, \beta_1, \beta_2)$ is an F-distribution with $df = (2n_2, 2n_1)$.

This follows from the results in EXAMPLE 4 and the result that the ratio of independent chisquare r.v.s/df has an $F-$distribution. It then follows that

$$1 - \alpha = P\left[F_{\frac{\alpha}{2}} \leq \frac{\left(2n_2\bar{Y}/\beta_2\right)/(2n_2)}{\left(2n_1\bar{X}/\beta_1\right)/(2n_1)} \leq F_{1-\frac{\alpha}{2}}\right],$$

where $F_{\frac{\alpha}{2}} = qf(\frac{\alpha}{2}, 2n_2, 2n_1)$ and $F_{1-\frac{\alpha}{2}} = qf(1-\frac{\alpha}{2}, 2n_2, 2n_1)$ are the lower and upper $\frac{\alpha}{2}$-percentiles of a F-distribution with $df = (2n_2, 2n_1)$.

Inverting the inequalities yields a $100(1-\alpha)\%$ C.I. for β_1/β_2 :

$$\left(\frac{\bar{X}}{\bar{Y}}F_{\frac{\alpha}{2}}, \frac{\bar{X}}{\bar{Y}}F_{1-\frac{\alpha}{2}}\right)$$

Note that in using the F-tables, the lower percentile is related to the upper percentile through $F_{\alpha,n_2,n_1} = 1/F_{1-\alpha,n_1,n_2}$,

$$\begin{aligned} 1 - \alpha = P[F_{n_2,n_1} \geq F_{\alpha,n_2,n_1}] &= P\left[\frac{1}{F_{n_2,n_1}} \leq \frac{1}{F_{\alpha,n_2,n_1}}\right] \\ &= \\ &= P\left[F_{n_1,n_2} \leq \frac{1}{F_{\alpha,n_2,n_1}}\right] \\ &= 1 - \alpha \end{aligned}$$

Therefore, we have that

$$F_{1-\alpha,n_1,n_2} = \frac{1}{F_{\alpha,n_2,n_1}}$$

EXAMPLE 6. C.I. for Population Proportion p

Suppose we have a population consisting of two types of units A or B with p being the proportion of Type A units. Alternatively, we may have a process that produces one of two types of units A or B with p being the probability of a Type A unit occurring. Let Y be the number of Type A outcomes in n *iid* trials or the number of Type A units occurring in a random sample taken with replacement from a population. In either case, $\hat{p} = Y/n$ and Y has a $Bin(n, p)$ distribution. A C.I. for p can be constructed in a number of ways.

The Clopper-Pearson C.I. for p:

See page 434 and Exercise 9.21 in Cassela-Berger for mathematical details.

The Clopper-Pearson $100(1-\alpha)\%$ C.I. for p can be expressed as

$$\{p \mid P[B(n;p) \leq y] \geq \alpha/2\} \cap \{p \mid P[B(n;p) \geq y] \geq \alpha/2\}$$

where the observed value of Y is y.

The interval can alternatively be expressed as $(P_L,\ P_U)$ where the values of P_L and P_U are obtained from the following equations: Suppose we observe Y=y in the study, then solve for P_L and P_U

$$(1) \quad \sum_{k=y}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = \frac{\alpha}{2}$$

$$(2) \quad \sum_{k=0}^{y} \binom{n}{k} P_U^k (1-P_U)^{n-k} = \frac{\alpha}{2}$$

Determining Limits for Clopper-Pearson C.I.

Case 1: If $y = 0$ then

$P_L = 0$ and $P_U = 1 - (\frac{\alpha}{2})^{1/n}$

The justification of these solutions is as follows:

If $y = 0$ then (1) $\Rightarrow \sum_{k=0}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = 1$ unless $P_L = 0$

If $y = 0$ then (2) $\Rightarrow \sum_{k=0}^{0} \binom{n}{k} P_U^k (1-P_U)^{n-k} = (1-P_U)^n = \frac{\alpha}{2} \Rightarrow P_U = 1 - (\frac{\alpha}{2})^{1/n}$

Case 2: If $y = n$

then $P_L = (\frac{\alpha}{2})^{1/n}$ and $P_U = 1$.

The justification of these solutions is as follows:

If $y = n$ then (2) $\Rightarrow \sum_{k=0}^{n} \binom{n}{k} P_U^k (1-P_U)^{n-k} = 1$ unless $P_U = 1$

If $y = n$ then (1) $\Rightarrow \sum_{k=n}^{n} \binom{n}{k} P_L^k (1-P_L)^{n-k} = P_L^n = \frac{\alpha}{2} \Rightarrow P_L = (\frac{\alpha}{2})^{1/n}$

Case 3: For $y = 1, 2, \cdots, n-1$; (more general case)

Using the relationship between the binomial distribution, beta distribution, and the Fisher - F distribution, the solution to the equations can be expressed by using the upper $\frac{\alpha}{2}$ percentiles from the

F distribution: $F_{df_1, df_2, \frac{\alpha}{2}} = qf(1 - \frac{\alpha}{2}, df_1, df_2)$:

$$P_L = \frac{1}{1 + \left(\frac{n-y+1}{y}\right) F_{2(n-y+1),2y,\frac{\alpha}{2}}}; \qquad P_U = \frac{\left(\frac{y+1}{n-y}\right) F_{2(y+1),2(n-y),\frac{\alpha}{2}}}{1 + \left(\frac{y+1}{n-y}\right) F_{2(y+1),2(n-y),\frac{\alpha}{2}}}$$

Examples when $y = 0$:

With $n = 20$, $y = 0$ then $\hat{p} = 0$ and 95% C.I. for p is

$P_L = 0$ and $P_U = 1 - (.025)^{1/20} = .168 \quad \Rightarrow (0, .168)$

With $n = 100$, $y = 0$ then $\hat{p} = 0$ and 95% C.I. for p is

$P_L = 0$ and $P_U = 1 - (.025)^{1/100} = .0362 \quad \Rightarrow (0, .0362)$

Examples when $y = n$:

With $n = 20$, $y = 20$ then $\hat{p} = 1$ and 95% C.I.. for p is

$P_L = (.025)^{1/20} = .832$ and $P_U = 1 \quad \Rightarrow (.832, 1)$

With $n = 100$, $y = 100$ then $\hat{p} = 1$ and 95% C.I. for p is

$P_L = (.025)^{1/100} = .9638$ and $P_U = 1 \quad \Rightarrow (.9638, 1)$

Example when $0 < y < n$ Suppose $n = 20$ and $y = 5$. Obtain a 95% C.I. for p.

$$F_{2(n-y+1),2y,\frac{\alpha}{2}} = F_{32,10,.025} = qf(1 - .025, 32, 10) = 3.297234$$

$$F_{2(y+1),2(n-y),\frac{\alpha}{2}} = F_{12,30,.025} = qf(1 - .025, 12, 30) = 2.412034$$

$$P_L = \frac{1}{1 + \left(\frac{16}{5}\right) F_{32,10,.025}} = \frac{1}{1 + \left(\frac{16}{5}\right)(3.297234)} = .087$$

$$P_U = \frac{\left(\frac{6}{15}\right) F_{12,30,.025}}{1 + \left(\frac{6}{15}\right) F_{12,30,.025}} = \frac{\left(\frac{6}{15}\right)(2.412034)}{1 + \left(\frac{6}{15}\right)(2.412034)} = .491$$

Therefore, the 95% C.I. for p is (.087, .491).

Comments

- What is the connection between the binomial distribution and the F-distribution?

  Suppose X is distributed $Bin(n, p)$, then

  $P[X \geq x] = P[Y \leq p]$ where Y is distributed $Beta(\alpha = x, \beta = n - x + 1)$.

  Furthermore, suppose W has an $F-$distribution with $df_1 = \nu_1, df_2 = \nu_2$ then

  $$\frac{\left(\frac{\nu_1}{\nu_2}\right) W}{1 + \left(\frac{\nu_1}{\nu_2}\right) W} \quad \text{has a} \quad Beta\left(\frac{\nu_1}{2}, \frac{\nu_2}{2}\right) \quad \text{distribution}$$

- One problem with the Clopper-Pearson C.I. is that it is necessarily conservative, that is, it has coverage probability greater than $1 - \alpha$. This is due to that the binomial distribution being discrete and hence it is impossible to obtain bounds having exactly a $1 - \alpha$ probability.

  Therefore, the bounds are set so that the coverage probability is at least $1-\alpha$. A problem with having a slightly higher coverage probability is that the width of the C.I. may be somewhat wider than a corresponding interval having a coverage probability of exactly $1 - \alpha$.

STOP Friday 10/22/21

START Monday 10/25/21

Asymptotic (large n) Results for the Sampling Distribution of Pivot

When we are unable to derive the exact sampling distribution of the pivot it may be able to obtain asymptotic (large n) approximations. For example, in EX 1, EX 2, and EX 3, even if the population distributions were non-normal but the sample sizes were large it would be possible to construct the C.I.'s using the central limit theorem results with the sampling distribution of the sample mean having approximately a normal distribution and S being a consistent estimator of σ for large n. Because n is large the t-based percentiles would be essentially the same as the standard normal percentiles. Therefore, the C.I.'s would have approximately the correct level of confidence. However, the sample size necessary to invoke the central limit theorem results varies depending on the true population distribution as was seen in Handout 10, sampling distribution handout. The following example will illustrate the problems that may result when n is too small.

Improper Use of Central Limit Theorem

If we have a simple random sample of size $n = 20$ from a $N(\mu, \sigma^2)$ distribution and construct 100 95% C.I.'s for μ using the t-distribution based pivot,

$$\bar{X} \pm t_{\frac{\alpha}{2},n-1} \frac{S}{\sqrt{n}}$$

we would expect approximately 5 of the 100 intervals to fail to contain μ. The fact that the t-distribution is the valid sampling distribution for the pivot,

$$\bar{X} \pm t_{\alpha/2,n-1} \frac{S}{\sqrt{n}}$$

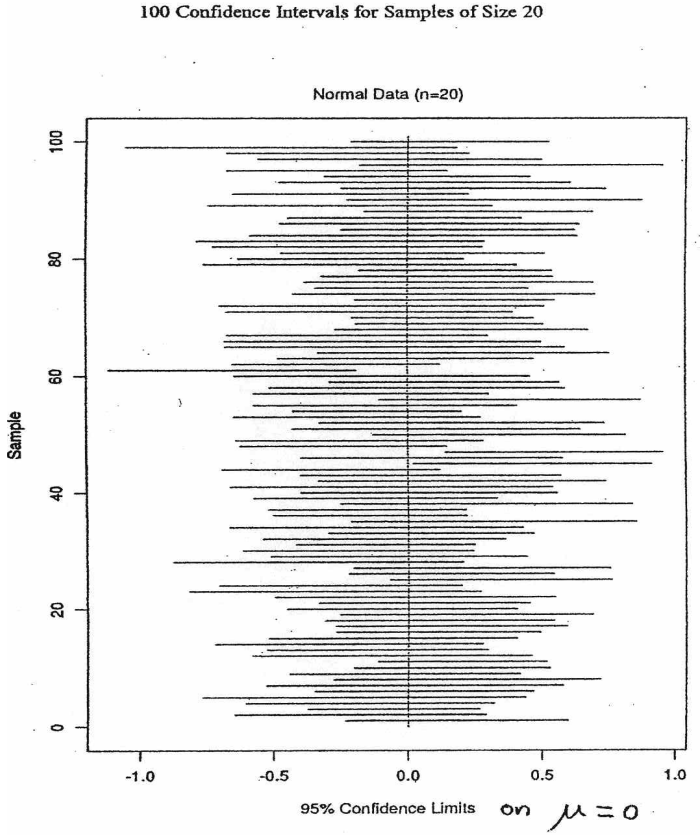
requires that

1. $\bar{X}$ have a normal distribution
2. $\frac{(n-1)S^2}{\sigma^2}$ have a chi-square distribution
3. $\bar{X}$ and S be independent.

These conditions are met when sampling from a $N(\mu, \sigma^2)$ distribution. This is illustrated in the graphs on the next page. Note that only 3 of the 100 C.I.'s fail to contain μ, the sampling distribution of $\bar{X}$ is nearly normal in shape and the scatterplot indicates that $\bar{X}$ and S are uncorrelated.

Next, suppose we have a simple random sample of size $n = 20$ from a $Exp(\beta)$ distribution and construct 100 95% C.I.'s for $\mu = \beta$ using the t-distribution. What is the impact of the skewness of the exponential distribution on the level of the C.I.'s. Note that now instead of having approximately 5 of the 100 C.I.'s fail to contain μ we now have 9 of the 100 C.I.'s fail to contain μ. Why is the coverage probability so much less than the stated 95%? Examining the graphs, we note that the sampling distribution of $\bar{X}$ appears normal in shape and the sampling distribution of S is similar to the shape from sampling from a normal distribution. However, the scatterplot of S vs $\bar{X}$ show a very strong positive correlation which would violate our independence requirement. Why are $\bar{X}$ and S correlated when sampling from an exponential distribution? A heuristic explanation is that in the exponential distribution both $\bar{X}$ and S are estimating the same parameter β because $\mu = \beta$ and $\sigma = \beta$. However, a similar sort of behavior can be seen in many other right skewed distributions.

Population Distribution is $N(0,1)$



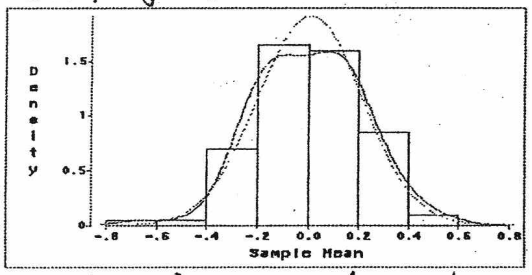
Based on $X_1, \cdots, X_{20}$ iid $N(0,1)$

$\bar{X} \pm t_{.025,19} \frac{S}{\sqrt{20}}$

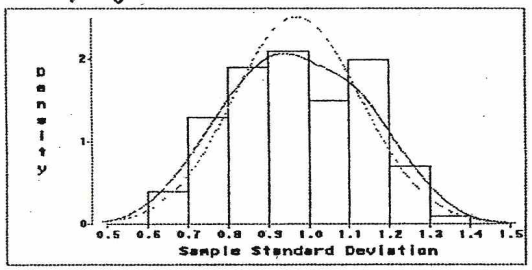
(100 Reps of the above Process)

Population Has a $N(0,1)$ Distribution

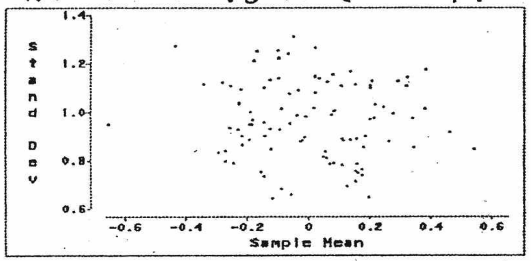
Sampling Distribution of $\bar{X}$ with $n=20$



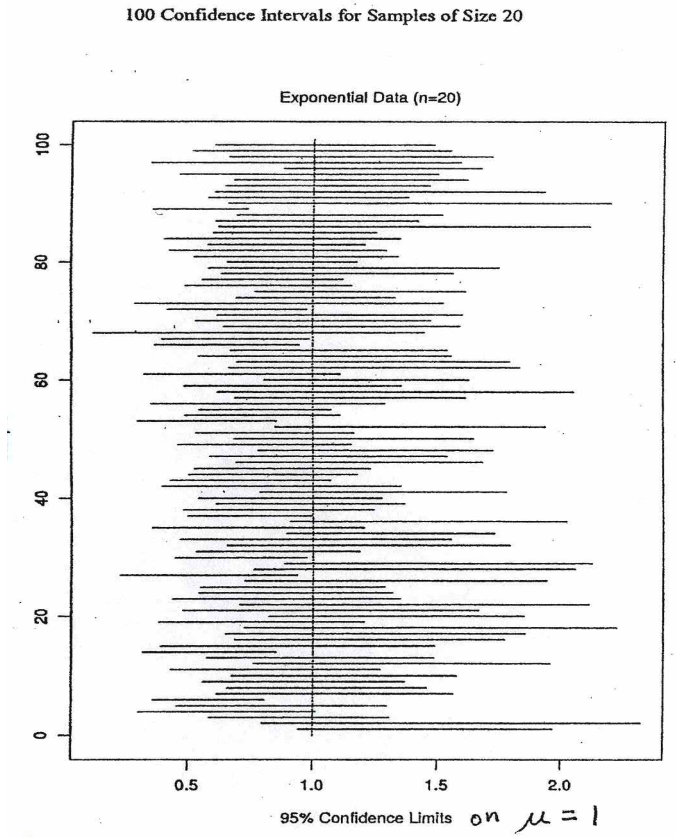
Sampling Distribution of S with $n=20$



Plot of S vs $\bar{X}$ (100 reps)



Population Distribution is $Exp(\lambda=1)$

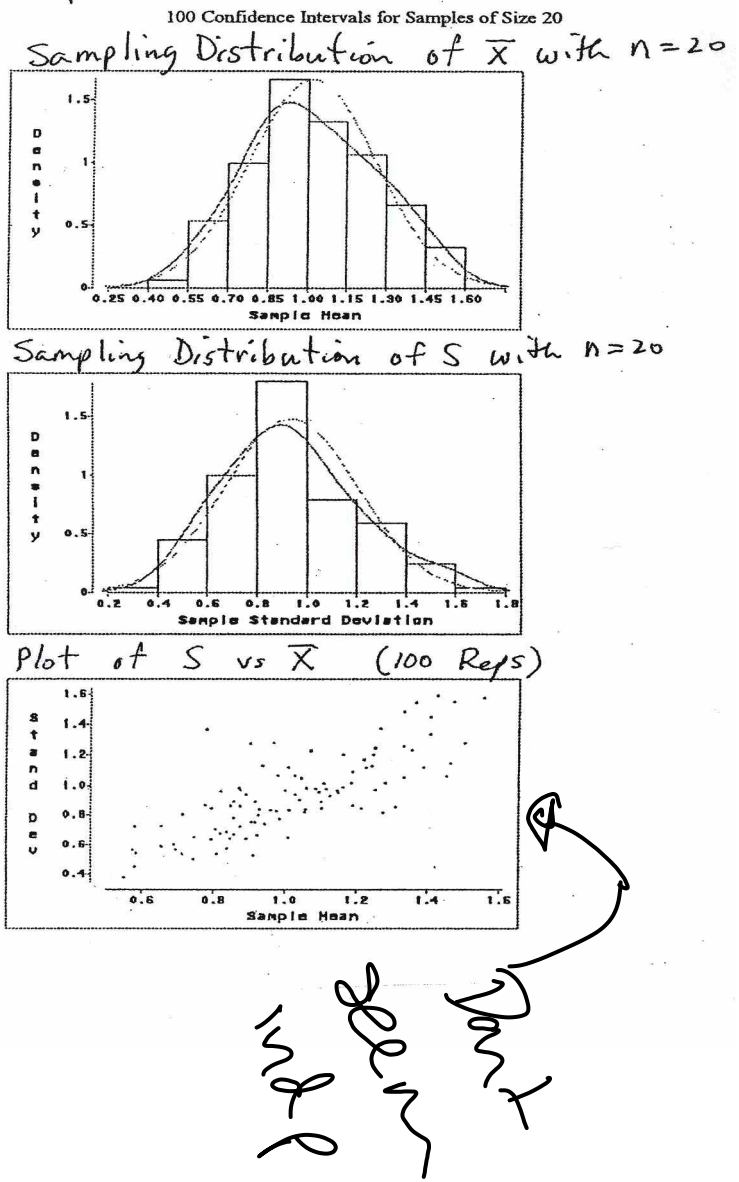


Based on $X_1, \cdots, X_{20}$ iid $Exp(\lambda=1)$

$\bar{X} \pm t_{.025} \frac{S}{\sqrt{20}}$

(100 Reps of the Above Prouss)

Population Distribution is $Exp(\lambda=1)$



Don't Memorize Me!

With very heavy tailed distributions, the sample standard deviation S has a much more skewed distribution than its distribution when sampling from a normal distribution. Thus the values of S are unusually small too often with a few very large values. With highly skewed distributions, the values of $\bar{Y}$ and S may be positively correlated when using a relatively small value for n. The following simulation results demonstrate the effects of these two results if the t-based 95% C.I. is constructed for a population mean μ: $\bar{Y} \pm t_{.025}S/\sqrt{n}$. Five symmetric distributions were used and their corresponding "half" distribution was obtained by folding the distribution over its point of symmetry. For each of the three samples sizes $n = 10, 20, 30$ and ten distributions, 10000 samples of size n were generated. The table provides the Coverage Probability of these 10000 95% C.I.'s for μ and the correlation between $\bar{Y}$ and S. (These results are from Dr. Cline).

n=10		
Distribution	Coverage Probability	Corr(Y, S)
Parabola	0.942	0.019
Half-Parabola	0.944	0.331
Normal	0.950	-0.003
Half-Normal	0.937	0.595
Double-Exponential	0.937	-0.006
Exponential	0.900	0.759
Symmetric Lognormal	0.960	0.038
Lognormal	0.839	0.866
Symmetric Pareto	0.839	-0.012
Pareto	0.839	0.853
n=20		
Distribution	Coverage Probability	Corr(Y, S)
Parabola	0.950	0.018
Half-Parabola	0.948	0.342
Normal	0.948	0.001
Half-Normal	0.946	0.590
Double-Exponential	0.946	-0.012
Exponential	0.918	0.750
Symmetric Lognormal	0.959	-0.027
Lognormal	0.868	0.853
Symmetric Pareto	0.868	-0.020
Pareto	0.868	0.817
n=30		
Distribution	Coverage Probability	Corr$(\bar{Y}, S)$
Parabola	0.951	0.020
Half-Parabola	0.951	0.349
Normal	0.948	-0.008
Half-Normal	0.941	0.600
Double-Exponential	0.941	0.009
Exponential	0.921	0.739
Symmetric Lognormal	0.960	0.007
Lognormal	0.884	0.825
Symmetric Pareto	0.884	-0.044
Pareto	0.884	0.789

If data are not normal 95% CIs do not actually contain 0 95% of the time.

Asymptotic C.I. for Population Proportion p - Wald Confidence Intervals

Let Y be the number of Type A outcomes in n *iid* trials or the number of Type A units occurring in a random sample taken with replacement from a population.

In either case, $\hat{p} = Y/n$ and Y has a $Bin(n, p)$ distribution.

Using the C.L.Th. for the binomial distribution we obtain the following results:

- The sampling distribution of $\hat{p} = Y/n$ has

  asymptotic mean and standard deviation: $\mu_A = p$ and $\sigma_A = \sqrt{p(1-p)}/\sqrt{n}$

  Therefore, the appropriate pivot is given by

- Pivot=$g(Y, p) = \frac{\hat{p}-p}{\sqrt{p(1-p)}/\sqrt{n}}$

  which has approximately a $N(0, 1)$ distribution for large n by the Central Limit Theorem.

Approach 1: Wald C.I. for p:

Replace p with $\hat{p}$ in the denominator of the pivot

$$g(Y, p) = \frac{\hat{p} - p}{\sqrt{\hat{p}(1-\hat{p})}/\sqrt{n}}$$

and require that $min(n\hat{p}, n(1-\hat{p})) \geq 5$.

This results in the Wald $100(1-\alpha)\%$ C.I. for p:

$$CI_S = \quad \hat{p} \pm Z_{\frac{\alpha}{2}} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}$$

The Wald C.I. does not provide adequate coverage for small n due to the estimation of the standard deviation of $\hat{p}$.

Also, the C.I. is meaningless if $\hat{p} = 0$ or $\hat{p} = 1$ because in both cases C.I. $= \hat{p} \pm 0$

Therefore, an alternative approach was given by Wilson.

Approach 2, Wilson C.I. for p

Wilson(1927) used the original pivot, with p not $\hat{p}$ to obtain a C.I. for p by inverting the following inequalities:

$$1-\alpha \approx P\left[\frac{|\hat{p}-p|}{\sqrt{p(1-p)}/\sqrt{n}} \leq Z_{\frac{\alpha}{2}}\right] \Rightarrow \text{ C.I. is } \frac{|\hat{p}-p|}{\sqrt{p(1-p)}/\sqrt{n}} \leq Z_{\frac{\alpha}{2}}$$

Squaring this interval yields

$$(\hat{p}-p)^2 \leq Z^2_{\frac{\alpha}{2}} \frac{p(1-p)}{n} \Rightarrow$$

$$h(p) = (1+C)p^2 - (C+2\hat{p})p + \hat{p}^2 \leq 0 \text{ where } C = \frac{1}{n} Z^2_{\frac{\alpha}{2}}$$

Next, need to solve inequality $h(p) \leq 0$ for p, that is, find the region $\{p : h(p) \leq 0\}$.

This yields the following region:

$$\frac{n\hat{p} + .5Z^2_{\frac{\alpha}{2}}}{n + Z^2_{\frac{\alpha}{2}}} \pm \frac{\sqrt{n} Z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{\frac{\alpha}{2}}}}{n + Z^2_{\frac{\alpha}{2}}}$$

Set $\tilde{Y} = Y + .5Z^2_{\frac{\alpha}{2}}, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \tilde{Y}/\tilde{n}$ yields the Wilson C.I. for p:

$$CI_W = \quad \tilde{p} \pm \frac{\sqrt{n} Z_{\frac{\alpha}{2}} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{\frac{\alpha}{2}}}}{\tilde{n}}$$

Note that for large n ($n \geq 40$),

$$\begin{aligned} \tilde{p} &= \frac{Y + .5Z^2_{\alpha/2}}{n + Z^2_{\alpha/2}} \\ &= \frac{Y}{n + Z^2_{\alpha/2}} + \frac{.5Z^2_{\alpha/2}}{n + Z^2_{\alpha/2}} \\ &\approx \frac{Y}{n} \\ &= \hat{p} \end{aligned}$$

This leads to the Agresti-Coull C.I. for p.

Approach 3, Agresti-Coull C.I. for p

Agresti-Coull C.I. for p:

$$CI_{AC} = \quad \tilde{p} \pm Z_{\frac{\alpha}{2}} \frac{\sqrt{\tilde{p}(1-\tilde{p})}}{\sqrt{\tilde{n}}}$$

Special Case of Agresti-Coull for 95% C.I.

For a 95% C.I. we have

$$Z_{\frac{\alpha}{2}} = Z_{.025} = 1.96 \approx 2 \Rightarrow$$

$$\tilde{Y} = Y + .5Z^2_{\frac{\alpha}{2}} \approx Y + 2 \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}} \approx n + 4 \quad \Rightarrow$$

$$\tilde{p} = \frac{Y+2}{n+4} \quad \text{Requires a level of 95\%}$$

This is the formula seen in many textbooks.

There are many other approaches but we will only consider four of these confidence intervals for p:

1. the confidence interval using binomial distribution, Clopper-Pearson interval: CI_{CP},
2. the Wald confidence interval using the asymptotic distribution: CI_{WALD},
3. the Wilson confidence interval: CI_{WILS}
4. the Agresti-Coull confidence interval: CI_{AC}.

The following table computes all four C.I.'s for a variety of values for n and Y to illustrate how the width of the intervals vary considerably from the Binomial based C.I.

When $Y = n\hat{p} < 5$, the Wald C.I.'s are not recommended.

The C.I.'s are presented to just illustrate their inaccuracies in these situations relative to the Binomial based C.I.'s.

From the formulas for the Wilson and Agresti-Coull C.I.'s we know that the Agresti-Coull C.I. is always wider than the Wilson C.I.

From the table, for n relatively small both Wilson and Agresti-Coull C.I.'s may be too narrow and hence have coverage probability less than 95%.

Comparison of Various 95% C.I.'s for Proportion

n	$\hat{p}$	y	Wald	Wilson	Agresti-Coull	Clopper-Pearson
10	.10	1	(.0000, .2859)	(.0179, .4042)	(.0000, .4260)	(.003, .445)
10	.20	2	(.0000, .4479)	(.0567, .5098)	(.0459, .5206)	(.025, .556)
10	.50	5	(.1901, .8099)	(.2366, .7634)	(.2366, .7634)	(.187, .813)
25	.04	1	(.0000, .1168)	(.0071, .1954)	(.0000, .2114)	(.001, .204)
25	.20	5	(.0432, .3568)	(.0886, .3913)	(.0841, .3958)	(.068, .407)
50	.02	1	(.0000, .0588)	(.0035, .1050)	(.0000, .1148)	(.001, .107)
50	.04	2	(.0000, .0943)	(.0110, .1346)	(.0034, .1422)	(.005, .137)
50	.10	5	(.0178, .1832)	(.0435, .2136)	(.0391, .2179)	(.003, .218)
50	.20	10	(.0891, .3109)	(.1124, .3304)	(.1105, .3323)	(.100, .338)
50	.50	25	(.3614, .6386)	(.3664, .6336)	(.3664, .6336)	(.355, .645)
100	.02	2	(.0000, .0474)	(.0055, .0700)	(.0011, .0744)	(.002, .070)
100	.04	4	(.0016, .0784)	(.0157, .0984)	(.0124, .1016)	(.011, .099)
100	.10	10	(.0412, .1588)	(.0552, .1744)	(.0535, .1761)	(.049, .176)
100	.20	20	(.1216, .2784)	(.1334, .2888)	(.1326, .2896)	(.127, .292)
100	.50	50	(.4020, .5980)	(.4038, .5962)	(.4038, .5962)	(.398, .602)
250	.02	5	(.0026, .0374)	(.0085, .0460)	(.0072, .0473)	(.007, .046)
250	.04	10	(.0157, .0643)	(.0218, .0721)	(.0209, .0730)	(.019, .072)
250	.10	25	(.0628, .1372)	(.0686, .1435)	(.0682, .1439)	(.066, .144)
250	.20	50	(.1504, .2496)	(.1551, .2540)	(.1549, .2542)	(.152, .255)
250	.50	125	(.4380, .5620)	(.4384, .5615)	(.4385, .5615)	(.436, .564)
500	.02	10	(.0077, .0323)	(.0108, .0364)	(.0104, .0369)	(.010, .036)
500	.04	20	(.0228, .0572)	(.0260, .0610)	(.0257, .0613)	(.025, .061)
500	.10	50	(.0737, .1263)	(.0766, .1294)	(.0765, .1296)	(.075, .130)
500	.20	100	(.1649, .2351)	(.1672, .2373)	(.1672, .2374)	(.161, .238)
500	.50	250	(.4562, .5438)	(.4563, .5437)	(.4563, .5437)	(.455, .545)
1000	.02	20	(.0113, .0287)	(.0129, .0307)	(.0128, .0309)	(.012, .031)
1000	.04	40	(.0279, .0521)	(.0295, .0540)	(.0294, .0541)	(.029, .054)
1000	.10	100	(.0814, .1186)	(.0829, .1201)	(.0828, .1202)	(.082, .120)
1000	.20	200	(.1752, .2248)	(.1763, .2259)	(.1764, .2259)	(.176, .226)
1000	.50	250	(.4690, .5310)	(.4690, .5309)	(.4691, .5309)	(.469, .531)

Note: When $y = n\hat{p} < 5$, the Asymptotic C.I.'s are not recommended.

The C.I.'s are given to illustrate their inaccuracies in these situations relative to the Clopper-Pearson C.I.

The Clopper-Pearson C.I.'s tend to produce intervals having coverage probabilities somewhat larger than the nominal value. This is due to the discreteness of the binomial distribution.

=> Clopper CI's tend to be conservative estimates.

- would prefer to be conservative

Comparison of Performance of C.I.'s

Given an interval estimator of a parameter θ: $100(1-\alpha)$ C.I. $= (\hat{\theta}_L, \hat{\theta}_U)$.

There a number of methods for assessing the performance of an $100(1-\alpha)$ C.I. as an estimator of θ:

1. Accuracy of C.I. is measured by Coverage Probability: $C(\theta, n) = P[\theta \epsilon (\hat{\theta}_L, \hat{\theta}_U)]$

   Compare $C(\theta, n)$ to $100(1-\alpha)$ to determine how close the true level of confidence is to the nominal (stated) level.

2. Precision of C.I. is measured by Expected Width of C.I.: $E[W(\theta, n)]$

   Because $(\hat{\theta}_L, \hat{\theta}_U)$ is a r.v., we need to compute its average width.

   That is, let $W(\theta, n) = \hat{\theta}_U - \hat{\theta}_L$ and then compute $E[W(\theta, n)]$.

   In comparing two C.I.'s for the same parameter having the same **coverage probability**, the C.I. having shortest expected width is the better C.I.

The article **Interval Estimation for a Binomial Proportion** in *Statistical Science*, Vol. 16, pp. 101-133, by L. Brown, T. Cai and A. DasGupta contains a discussion of the performance of various confidence intervals for p using the above measures and several others. I have included a few of the graphs from their article. The following recommendations are given in the article:

- For larger n, the Wilson and A-C are comparable.
- The A-C CI is recommended for $n \geq 40$ due to its ease of calculation.
- For $n < 40$, there are several alternatives:

  1. Jeffreys Confidence Interval: This is a Bayesian confidence interval involving a $B(n, p)$ data distribution with a beta prior distribution on p, $Beta(a_1, a_2)$ (See STAT 638 for details).

     Observe Y distributed $B(n, p)$ then a $100(1-\alpha)\%$ Bayesian confidence interval on p is

     $$[Beta(\alpha/2;\ Y + a_1,\ n - Y + a_2),\ Beta(1-\alpha/2;\ Y + a_1,\ n - Y + a_2)]$$

     where $Beta(\alpha;\ m_1,\ m_2)$ denotes the α quantile of a $Beta(m_1, m_2)$ distribution.

     The values of a_1 and a_2 depend on the researcher's prior knowledge of the value of p.

  2. Several other confidence intervals are discussed in the article by Agresti and Coull (1998), "Approximate is Better than Exact for Interval Estimation of Binomial Proportions", *The American Statistican*, **Vol. 52**.

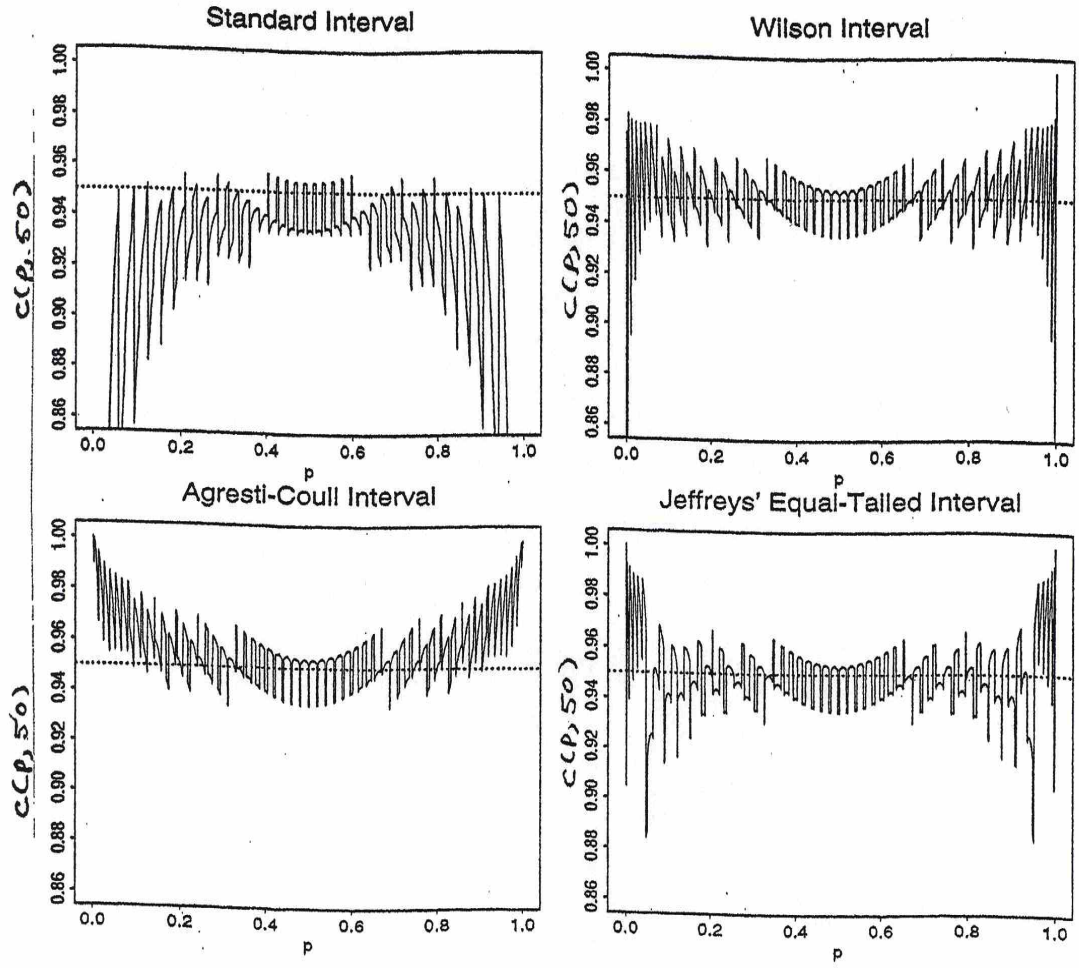


FIG. 5. *Coverage probability for* $n = 50$. (nominal = 0.95)

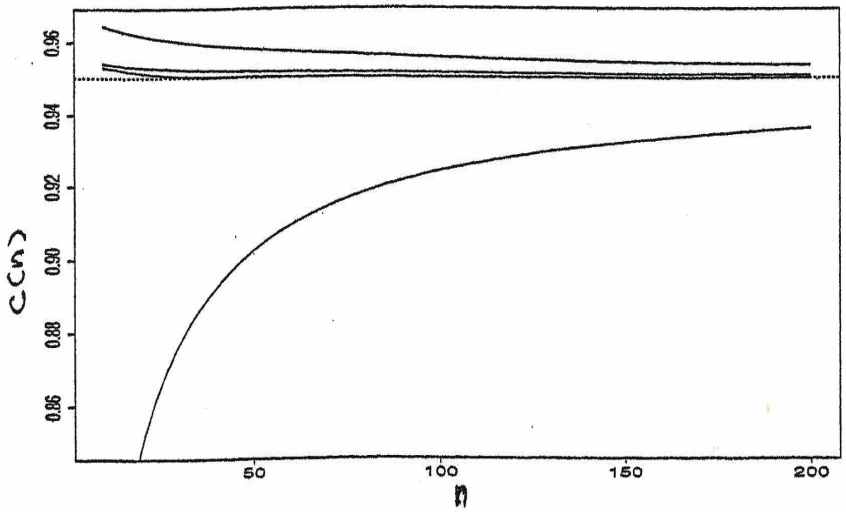


FIG. 6. *Comparison of the average coverage probabilities. From top to bottom: the Agresti–Coull interval* CI_{AC}, *the Wilson interval* CI_W, *the Jeffreys prior interval* CI_J *and the standard interval* CI_s. *The nominal confidence level is* 0.95. (Averaged over p)

3.3 Expected Length

Besides coverage, length is also very important in evaluation of a confidence interval. We compare both the expected length and the average expected length of the intervals. By definition,

$$\begin{aligned}&\text{Expected length}\\&\quad = E_{n,p}(\text{length}(CI))\\&\quad = \sum_{x=0}^{n}(U(x,n)-L(x,n))\binom{n}{x}p^x(1-p)^{n-x},\end{aligned}$$

where U and L are the upper and lower limits of the confidence interval CI, respectively. The average expected length is just the integral $\int_0^1 E_{n,p}(\text{length}(CI))\,dp$.

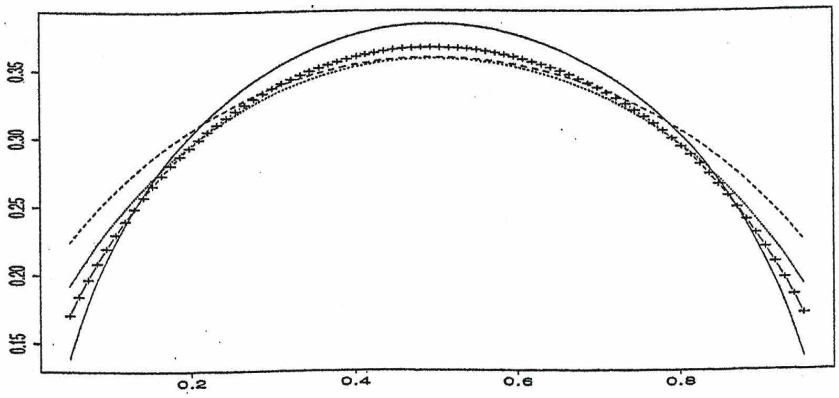


FIG. 8. *The expected lengths of the standard (solid), the Wilson (dotted), the Agresti–Coull (dashed) and the Jeffreys (+) intervals for $n = 25$ and $\alpha = 0.05$.*

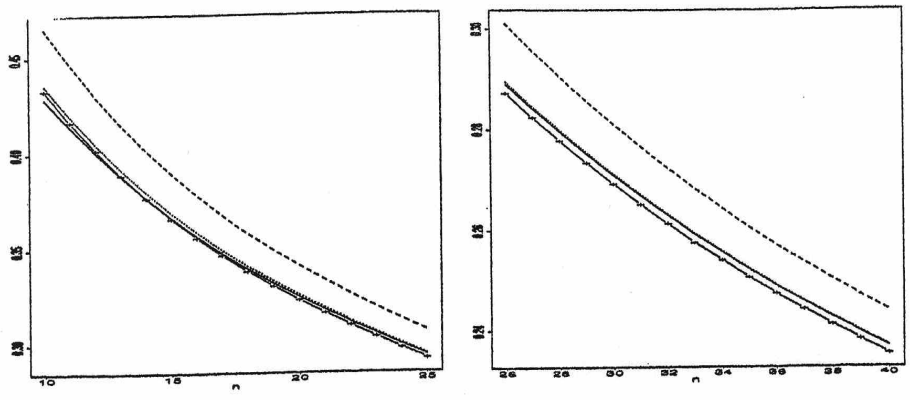


FIG. 9. *The average expected lengths of the standard (solid), the Wilson (dotted), the Agresti–Coull (dashed) and the Jeffreys (+) intervals for $n = 10$ to 25 and $n = 26$ to 40.*

Sample Size Determination

Statisticians are often asked the question "How much data do I need to collect?" One technique to answer this question is by way of C.I.'s.

Find the sample size n such that the estimator $\hat{\theta}$ of the parameter θ is within Δ units of the true value of θ with $100(1-\alpha)\%$ confidence.

We can then set up an equation and solve for the sample size n once the client provides the values of $100(1-\alpha)\%$ and Δ.

Consider the following two situations:

1. Sample Size for Estimating Population Mean μ

   Find n such that we are $100(1-\alpha)\%$ confident that $\bar{Y}$ is within Δ units of μ. The asymptotic sampling distribution for $\bar{Y}$ yields

   $$P\left[|\bar{Y}-\mu| \leq \Delta\right] = P\left[\left|\frac{\bar{Y}-\mu}{\sigma/\sqrt{n}}\right| \leq \Delta\frac{\sqrt{n}}{\sigma}\right] \approx 1-\alpha.$$

   Thus, set $\Delta\frac{\sqrt{n}}{\sigma} = Z_{\frac{\alpha}{2}}$ and solve for n yielding

   $$n = \frac{Z^2_{\frac{\alpha}{2}}\sigma^2}{\Delta^2}.$$

   Note that σ is often unknown so the client will have to have at least a rough guess of this value.

   This can be obtained from previous studies, literature, pilot study, or by using the crude estimator

   $$\hat{\sigma} \approx \frac{Range}{4}.$$

2. Sample Size for Estimating Population Proportion p

   Find n such that we are $100(1-\alpha)\%$ confident that $\hat{p}$ is within Δ units of p. The asymptotic sampling distribution for $\hat{p}$ yields

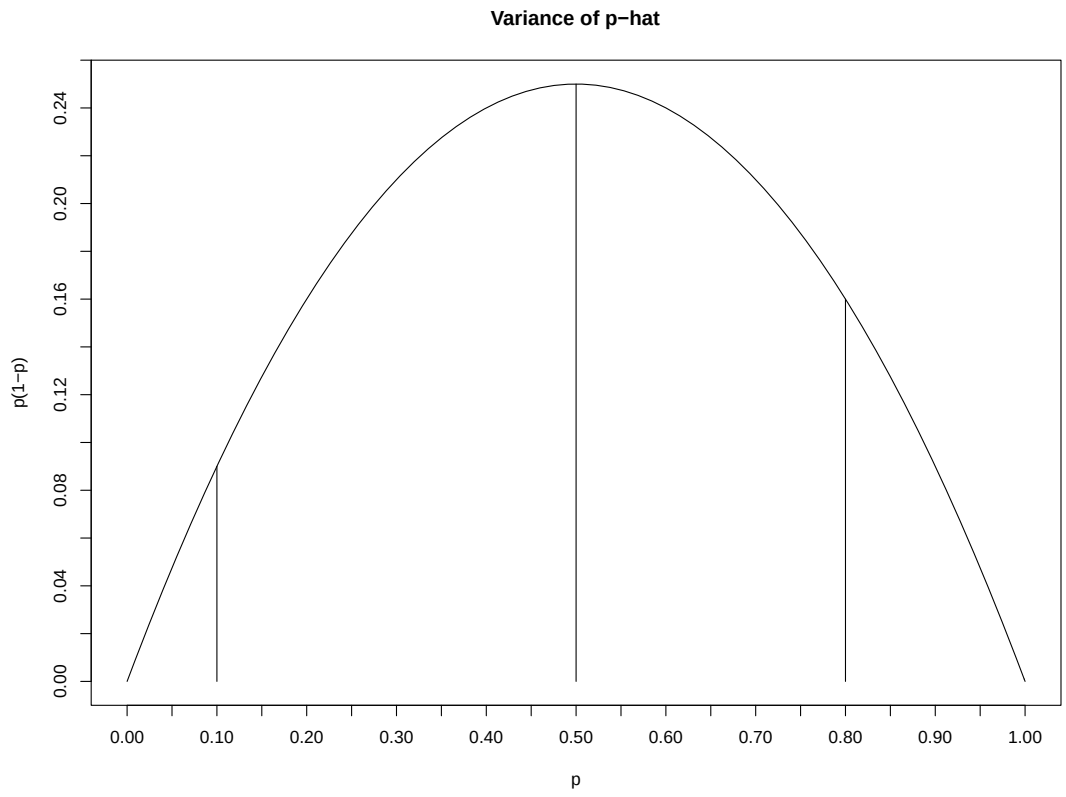
   $$P\left[|\hat{p}-p| \leq \Delta\right] = P\left[\left|\frac{\hat{p}-p}{\sqrt{p(1-p)}/\sqrt{n}}\right| \leq \Delta\frac{\sqrt{n}}{\sqrt{p(1-p)}}\right] \approx 1-\alpha.$$

   Thus, set $\Delta\frac{\sqrt{n}}{\sqrt{p(1-p}} = Z_{\frac{\alpha}{2}}$ and solve for n yielding

   $$n = \frac{Z^2_{\frac{\alpha}{2}}p(1-p)}{\Delta^2}.$$

   Note that p, an unknown value, is in the formula so the client will have to have at least a rough guess of the value of p.

When the client can provide a bound on p by knowing that $p \leq p_L$ or $p \geq p_U$.



Then replace p with either p_L or p_U.

If you cannot bound p away from 0.5 then just replace p with 0.5

Replacing p with 0.5 generally will produce very large values for n.

Example Find n such that 99% confident that $\hat{p}$ is within 0.01 units of p.

a. Suppose we know that $0 \leq p \leq 0.10$

   Then, $n = \frac{(2.576)^2(.1)(.9)}{(.01)^2} = 5973.$

b. Suppose we know that $.8 \leq p \leq 1.0$

   Then, $n = \frac{(2.576)^2(.8)(.2)}{(.01)^2} = 10618.$

c. If we were not able to bound p away from 0.5 then

   $n = \frac{(2.576)^2(.5)(.5)}{(.01)^2} = 16,590.$

Conservative way to choose n is to choose p closest to 0.5

3. More accurate calculations can be obtained using the formula for the Wilson CI:

$$n = \frac{2z^2_{\frac{\alpha}{2}}\hat{p}\hat{q} - 4z^2_{\frac{\alpha}{2}}\Delta^2 + \sqrt{4z^4_{\frac{\alpha}{2}}\hat{p}\hat{q}(\hat{p}\hat{q} - 4\Delta^2) + 4\Delta^2 z^4_{\frac{\alpha}{2}}}}{4\Delta^2}$$

For example, find n such that 99% confident that $\hat{p}$ is within 0.01 units of p.

Suppose we know that $0 \leq p \leq 0.10$. Then,

$$n = \frac{2(2.576)^2(.1)(.9) - 4(2.576)^2(.01)^2 + \sqrt{4(2.576)^4(.1)(.9)((.1)(.9) - 4(.01)^2) + 4(.01)^2(2.576)^4}}{4(.01)^2}$$

$$n = 5977.3 \Rightarrow$$

$$n = 5978$$

Distribution-Free C.I. for Population Quantile Q(u)

Suppose $Y_1, \ldots, Y_n$ are iid with strictly increasing continuous cdf $F(\cdot)$ and quantile function $Q(\cdot) = F^{-1}(\cdot)$.

A $100(1-\alpha)$ C.I. for $Q(u)$ is

$$\left(Y_{(r)},\ Y_{(s)}\right)$$

where $Y_{(1)} < \cdots < Y_{(n)}$, and r is the largest integer and s is the smallest integer such that

$$1 \le r < s \le n \ \text{ and } \ P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] \ge 1-\alpha$$

The values of r and s are selected such that

$$1-\alpha = \sum_{j=r}^{s-1} \binom{n}{j} u^j (1-u)^{n-j} = P\left[r \le B \le s-1\right]$$

where B is $Bin(n,u)$.

This result is obtained from

$$\begin{aligned} P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] &= P\left[F(Y_{(r)}) \le F(Q(u)) \le F(Y_{(s)})\right] \\ &= P\left[U_{(r)} \le u \le U_{(s)}\right] \end{aligned}$$

$F(Y) \sim Unif(0,1)$

where $U_{(1)} < \cdots < U_{(n)}$ are order statistics from iid Uniform on (0,1) distribution.

Next, we observe that

$$P\left[U_{(r)} \le u \le U_{(s)}\right] = P\left[\text{at least } \ n-s+1 \ U_{i's} \ge u \ \text{ and at least } \ r \ U_{i's} \le u\right] \ \Rightarrow$$

$$P\left[U_{(r)} \le u \le U_{(s)}\right] = P\left[\text{at most } \ s-1 \ U_{i's} \le u \ \text{ and at least } \ r \ U_{i's} \le u\right] \ \Rightarrow$$

$$\begin{aligned} P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] &= P\left[U_{(r)} \le u \le U_{(s)}\right] \\ &= \sum_{j=r}^{s-1} \binom{n}{j} (P(U_i \le u)^j (1-P(U_i \le u))^{n-j} \\ &= \sum_{j=r}^{s-1} \binom{n}{j} u^j (1-u)^{n-j} \\ &= P\left[r \le B \le s-1\right] = pbinom(s-1,n,u) - pbinom(r-1,n,u) \end{aligned}$$

where B is $Bin(n,u)$.

Finally, select the values of r and s such that

$$P\left[Y_{(r)} \le Q(u) \le Y_{(s)}\right] = P\left[r \le B \le s-1\right] = 1-\alpha$$

In most cases, we cannot obtain exactly, $1-\alpha$, in the above expressions so the coverage probability is always taken to be as close to $1-\alpha$ as possible but never less that $1-\alpha$.

STOP Monday 10/25/2021

EXAMPLE Suppose we wanted to find a 95% C.I. for the upper quartile,Q(.75), based on n=50 iid observations on the cdf F. The following Rcode will determine the values of r, s, and the true coverage:

```
n=50
L=.95
P=.75
s=ceiling(n*P)-1
r=floor(n*P)+1
cov=0
while(s<n-1 && r>1 && cov<L)
{s=s+1
cov=pbinom(s-1,n,P)-pbinom(r-1,n,P)
if(cov>=L) break;
r=r-1
cov=pbinom(s-1,n,P)-pbinom(r-1,n,P)
}
r
s
cov
> r
[1] 32
> s
[1] 44
> cov
[1] 0.951876
```

The 95% C.I. on Q(.75) would be $(Y_{(32)},\ Y_{(44)})$ with coverage probability of 95.2%.

Special Case: C.I. for Median

A C.I. for the population median is obtained by just setting

$u = .5$ and requiring that the C.I. to be symmetric, that is, $s = n - r + 1$.

This then yields a $100(1-\alpha)\%$ C.I. for the median $Q(.5)$:

$$\left(Y_{(r)},\ Y_{(n-r+1)}\right)$$

where r is the largest integer such that

$$1 - \alpha \leq P\left[r \leq B \leq n - r\right]$$

and B is $Bin(n, .5)$.

The following example will illustrate how to use R code to select the value of r and determine the true coverage probability.

Example Suppose we want a 95% C.I. on $Q(.5)$ based on $n = 50$.

Find the largest r such that $.95 \leq P\left[r \leq B \leq 50 - r\right]$, where B is $Bin(50, .5)$.

```
n =  50
cov =  .95
r =  0
imin =  0
```

```
i =  0
ans =  0
anst =  0
m =  1:n
ans =  pbinom(n-m,n,.5)-pbinom(m-1,n,.5)
while(i<n)
{
i =  i+1
if(ans[i]<cov)  anst[i] =  2
if(ans[i]>=cov) anst[i] =  ans[i]
}
ansmin =  min(anst)
imin =  which(anst==ansmin)
r =  imin
coverage =  ans[r]
```

From the above R-code we have $r = 18$ with coverage probability 0.96716.
Therefore, the 95% C.I. for the median is

$\left(Y_{(r)}, Y_{(n-r+1)}\right) = \left(Y_{(18)}, Y_{(33)}\right)$.

The true coverage probability is computed using the $B(50, .5)$ distribution.
Coverage Probability = (see previous page) $P\left[r \leq B \leq n-r\right] = P\left[18 \leq B \leq 32\right] = .96716$
So the true coverage probability is a little higher than 0.95.

The following table from *CRC Handbook of Tables for Probability and Statistics* provides r for a variety of values for n and levels of confidence.

START Wednesday 10/27/21

CRC Handbook of Tables for Probability and Statistics

VII.3 CONFIDENCE INTERVALS FOR MEDIANS

If the observations $x_1, x_2, \ldots, x_n$ are arranged in ascending order $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$, a $100(1-\alpha)\%$ confidence interval on the median of the population can be found. This table gives values of k and α such that one can be $100(1-\alpha)\%$ confident that the population median is between $x_{(k)}$ and $x_{(n-k+1)}$.

A—compute using wilcoxon signed rank test

CONFIDENCE INTERVALS FOR THE MEDIAN

n	Largest k	Actual $\alpha \leq 0.05$	Largest k	Actual $\alpha \leq 0.01$	N	Largest k	Actual $\alpha \leq 0.05$	Largest k	Actual $\alpha \leq 0.01$
6	1	0.031			36	12	0.029	10	0.004
7	1	0.016			37	13	0.047	11	0.008
8	1	0.008	1	0.008	38	13	0.034	11	0.005
9	2	0.039	1	0.004	39	13	0.024	12	0.009
10	2	0.021	1	0.002	40	14	0.038	12	0.006
11	2	0.012	1	0.001	41	14	0.028	12	0.004
12	3	0.039	2	0.006	42	15	0.044	13	0.008
13	3	0.022	2	0.003	43	15	0.032	13	0.005
14	3	0.013	2	0.002	44	16	0.049	14	0.010
15	4	0.035	3	0.007	45	16	0.036	14	0.007
16	4	0.021	3	0.004	46	16	0.026	14	0.005
17	5	0.049	3	0.002	47	17	0.040	15	0.008
18	5	0.031	4	0.008	48	17	0.029	15	0.006
19	5	0.019	4	0.004	49	18	0.044	16	0.009
20	6	0.041	4	0.003	50	18	0.033	16	0.007
21	6	0.027	5	0.007	51	19	0.049	16	0.005
22	6	0.017	5	0.004	52	19	0.036	17	0.008
23	7	0.035	5	0.003	53	19	0.027	17	0.005
24	7	0.023	6	0.007	54	20	0.040	18	0.009
25	8	0.043	6	0.004	55	20	0.030	18	0.006
26	8	0.029	7	0.009	56	21	0.044	18	0.005
27	8	0.019	7	0.006	57	21	0.033	19	0.008
28	9	0.036	7	0.004	58	22	0.048	19	0.005
29	9	0.024	8	0.008	59	22	0.036	20	0.009
30	10	0.043	8	0.005	60	22	0.027	20	0.006
31	10	0.029	8	0.003	61	23	0.040	21	0.010
32	10	0.020	9	0.007	62	23	0.030	21	0.007
33	11	0.035	9	0.005	63	24	0.043	21	0.005
34	11	0.024	10	0.009	64	24	0.033	22	0.008
35	12	0.041	10	0.006	65	25	0.046	22	0.006

The following table provides C.I.s for a variety of situations and parameters.

Parameter	Population Conditions	Endpoints of Confidence Intervals
Q(p)	$X_1, \cdots, X_n$ iid cont. cdf	$(X_{(r)}, X_{(s)})$,
		where r,s selected using Binomial(n,p) tables
μ	$X_1, \cdots, X_n$ iid $N(\mu, \sigma^2)$ σ unknown	$\bar{X} \pm t_{(\frac{\alpha}{2}, df)} \frac{S}{\sqrt{n}}$ where t has d.f.= n-1
$\mu_1 - \mu_2$	$X_1, \cdots, X_{n1}$ iid $N(\mu_1, \sigma_1^2)$ $Y_1, \cdots, Y_{n2}$ iid $N(\mu_2, \sigma_2^2)$ X's,Y's ind., $\sigma_1 = \sigma_2$	$(\bar{X} - \bar{Y}) \pm t_{(\frac{\alpha}{2}, df)} S_p \sqrt{\frac{1}{n1} + \frac{1}{n2}}$ where t has d.f.= $n1 + n2 - 2$
$\mu_1 - \mu_2$	$X_1, \cdots, X_{n1}$ iid $N(\mu_1, \sigma_1^2)$ $Y_1, \cdots, Y_{n2}$ iid $N(\mu_2, \sigma_2^2)$ X's,Y's ind., $\sigma_1 \neq \sigma_2$	$(\bar{X} - \bar{Y}) \pm t_{(\frac{\alpha}{2}, df)} \sqrt{\frac{S_1^2}{n1} + \frac{S_2^2}{n2}}$ where t has d.f.= $\frac{(C+1)^2 (n1-1)(n2-1)}{C^2(n2-1)+(n1-1)}$, and $C = \frac{S_1^2/n_1}{S_2^2/n_2}$
$\mu_1 - \mu_2$	$(X_1, Y_1), \cdots, (X_n, Y_n)$ iid with $D_i = X_i - Y_i$ $N(\mu_D, \sigma_D^2)$	$\bar{D} \pm t_{(\frac{\alpha}{2}, df)} S_D / \sqrt{n}$ where t has d.f. = n-1
p	Y is Bin(n,p)	$\tilde{p} \pm \frac{Z_{(\frac{\alpha}{2})} \sqrt{n} \sqrt{\hat{p}(1-\hat{p}) + \frac{1}{4n} Z^2_{(\frac{\alpha}{2})}}}{n + Z^2_{\frac{\alpha}{2}}}$
	$min(n\hat{p}, n(1-\hat{p}) \geq 5$	$Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
	and $n \leq 40$	$\tilde{Y} = Y + Z^2_{\frac{\alpha}{2}}/2, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \frac{\tilde{Y}}{\tilde{n}}$
p	Y is Bin(n,p)	$\tilde{p} \pm Z_{(\frac{\alpha}{2})} \sqrt{\tilde{p}(1-\tilde{p})/\tilde{n}}$
	$min(n\hat{p}, n(1-\hat{p}) \geq 5$	$Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
	and $n > 40$	$\tilde{Y} = Y + Z^2_{\frac{\alpha}{2}}/2, \quad \tilde{n} = n + Z^2_{\frac{\alpha}{2}}, \quad \tilde{p} = \frac{\tilde{Y}}{\tilde{n}}$
p	Y is Bin(n,p)	Clopper-Pearson CI: $(P_L, \ P_U)$, where $P_L = \frac{1}{1+\left(\frac{n-y+1}{y}\right) F_{2(n-y+1), 2y, \frac{\alpha}{2}}}$; $\quad P_U = \frac{\left(\frac{y+1}{n-y}\right) F_{2(y+1), 2(n-y), \frac{\alpha}{2}}}{1+\left(\frac{y+1}{n-y}\right) F_{2(y+1), 2(n-y), \frac{\alpha}{2}}}$ Upper $F-$ quantiles: $F_{df_1, df_2, \frac{\alpha}{2}} = qf(1 - \frac{\alpha}{2}, df_1, df_2)$

Parameter	Population Conditions	Endpoints of Confidence Intervals
$p_1 - p_2$	Count Data $min(n\hat{p}_i, n(1-\hat{p}_i) \geq 5$	$\hat{p}_1 - \hat{p}_2 \pm Z_{(\frac{\alpha}{2})}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$ $Z_{(\frac{\alpha}{2})}$ upper N(0,1) percentile
σ	Normal Data	$\left(\frac{\sqrt{n-1}\ \ S}{\sqrt{\chi^2_{(\frac{\alpha}{2},n-1)}}},\ \frac{\sqrt{n-1}\ \ S}{\sqrt{\chi^2_{(1-\frac{\alpha}{2},n-1)}}}\right)$, $\chi^2_{(\frac{\alpha}{2},n-1)}$ and $\chi^2_{(1-\frac{\alpha}{2},n-1)}$ upper percentiles- Chi-square tables
$\frac{\sigma_1}{\sigma_2}$	Normal Data	$\left(\frac{S_1}{S_2}\sqrt{\frac{1}{F_{(\frac{\alpha}{2},n_1-1,n_2-1)}}},\ \frac{S_1}{S_2}\sqrt{F_{(\frac{\alpha}{2},n_2-1,n_1-1}}\right)$, $F_{(\frac{\alpha}{2},n_1-1,n_2-1)}$ and $F_{(\frac{\alpha}{2},n_2-1,n_1-1)}$ upper percentiles- F-tables
β	Exponential Data	$\left(\frac{2n\bar{Y}}{\chi^2_{(1-\frac{\alpha}{2},2n)}},\ \frac{2n\bar{Y}}{\chi^2_{(\frac{\alpha}{2},2n)}}\right)$, $\chi^2_{(\frac{\alpha}{2},2n)}$ and $\chi^2_{(1-\frac{\alpha}{2},2n)}$ upper percentiles- Chi-square tables
$\frac{\beta_1}{\beta_2}$	Exponential Data	$\left(\frac{\bar{Y}_1}{\bar{Y}_2}\frac{1}{F_{(\frac{\alpha}{2},2n_1,2n_2)}},\ \frac{\bar{Y}_1}{\bar{Y}_2}F_{(\frac{\alpha}{2},2n_2,2n_1)}\right)$, $F_{(\frac{\alpha}{2},2n_1,2n_2)}$ and $F_{(\frac{\alpha}{2},2n_2,2n_1)}$ upper percentiles- F-tables
θ	parameter in pdf $f(y,\theta)$	$\hat{\theta} \pm Z_{\frac{\alpha}{2}}\widehat{SE}(\hat{\theta})$ where $\hat{\theta}$ is MLE

Prediction Intervals

In some studies or experiments, the researcher will want to predict the next outcome of the process or experiment.

For example, suppose your company is interested in purchasing a new airplane engine from supplier X. Your company is not interested in the average time to failure of all engines of that model but rather what is the predicted time to failure of the randomly selected engine that your company will be receiving.

Other examples, predicting the amount of rainfall in the next month, predicting demand for a product in the next quarter, and predicting the enrollment in a large undergraduate class for the next semester. Many of these types of predictions will use time series modeling and explanatory variables in regression models which you will study in STAT 626 and STAT 608.

An interval estimator of this predicted value is called a $100(1-\alpha)\%$ Prediction Interval P.I.

Case 1: Prediction Interval for a $N(\mu, \sigma^2)$ Population Distribution

Let $Y_1, \ldots, Y_n$ be iid $N(\mu, \sigma^2)$.

We can write the $Y_i's$ in the following model:

$$Y_i = \mu + \sigma Z_i \quad \text{where} \quad Z_i's \text{ are } iid \quad N(0,1)$$

The next unit selected from the population will have measured value:

$$Y_{n+1} = \mu + \sigma Z_{n+1}$$

Our estimator of Y_{n+1} is obtained by

1. Replacing μ and σ with their MLE's and
2. Replacing Z_{n+1} with $E[Z_{n+1}] = 0$, yielding

$$\hat{Y}_{n+1} = \hat{\mu} + \hat{\sigma}\hat{Z}_{n+1} = \hat{\mu} = \bar{Y}$$

To obtain, a P.I. for Y_{n+1} we need a pivot which involves the quantities

$$Y_{n+1}, \quad \hat{Y}_{n+1}, \quad \text{and} \quad \hat{\sigma}:$$

A possible candidate is

$$\text{Pivot } = g(Y, Y_{n+1}, \sigma) = \frac{Y_{n+1} - \bar{Y}}{S\sqrt{\frac{n+1}{n}}}$$

This will be a good candidate for a pivot only if we can determine its distribution.

From previous results, we know that the pivot has a t-Distribution with $df = n - 1$. Why?

1. $Y_{n+1} = \mu + \sigma Z_{n+1} \Rightarrow Y_{n+1}$ is distributed $N(\mu, \sigma^2)$
2. $Y_{n+1}, \bar{Y}$ are independent $\Rightarrow Y_{n+1} - \bar{Y}$ is distributed $N\left(\mu - \mu, \sigma^2 + \frac{\sigma^2}{n}\right) = N\left(0, \sigma^2\left(\frac{n+1}{n}\right)\right)$
3. $Y_{n+1} - \bar{Y}$ is distributed independent of S which has $\frac{(n-1)S^2}{\sigma^2}$ distributed as Chi-square with df=n-1
4. $t = \frac{N(0,1)}{\sqrt{\text{Chi-square/df}}} \Rightarrow \frac{(Y_{n+1}-\bar{Y})/\sigma\sqrt{\left(\frac{n+1}{n}\right)}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} = \frac{(Y_{n+1}-\bar{Y})}{S\sqrt{\left(\frac{n+1}{n}\right)}}$

Thus, we have the following results:

$$\begin{aligned} 1 - \alpha &= P\left[-t_{\frac{\alpha}{2}} \leq \frac{Y_{n+1} - \bar{Y}}{S\sqrt{\frac{n+1}{n}}} \leq t_{\frac{\alpha}{2}}\right] \\ &= P\left[\bar{Y} - t_{\frac{\alpha}{2}} S\sqrt{\frac{n+1}{n}} \leq Y_{n+1} \leq \bar{Y} + t_{\frac{\alpha}{2}} S\sqrt{\frac{n+1}{n}}\right] \end{aligned}$$

The $100(1-\alpha)\%$ Prediction Interval for Y_{n+1} is

$$\bar{Y} \pm t_{\frac{\alpha}{2}} S\sqrt{1 + \frac{1}{n}}$$

Recall that the $100(1-\alpha)\%$ C.I. for μ in a $N(\mu, \sigma^2)$ population is

$$\bar{Y} \pm t_{\frac{\alpha}{2}} S\sqrt{\frac{1}{n}}$$

Thus, the added width in the P.I. reflects the additional uncertainty in attempting to predict a realization from a $N(\mu, \sigma^2)$ population in comparison to estimating the population mean.

Case 2: Prediction Interval for a $Exponential(\beta)$ Distribution

Suppose the data $W_1, \ldots, W_n$ is iid $Exponential(\beta)$ r.v.'s with β unknown

A prediction is needed for the next realization from this population.

For example, the time to the next hurricane in the Gulf of Mexico or the life length of the transmission placed in the Porsche 911 Turbo that you will purchase upon graduation.

The MLE point estimator of β is $\hat{\beta} = \bar{W}$.

A pivot should involve

- the point estimator, $\hat{\beta} = \bar{W}$
- the next realization W_{n+1}, and
- the unknown parameter β.

  Recall that if $Y_1, \ldots, Y_k$ are iid $Exp(\beta)$ then

- $k\bar{Y} = \sum_{i=1}^{k} Y_i$ has a $Gamma(k, \beta)$ distribution.
- If X has a $Gamma(k, \beta)$ distribution, then $\frac{2X}{\beta}$ has a chi-square distribution with $df = 2k$.

  Therefore, we know have the following results:

1. $\frac{2W_{n+1}}{\beta}$ has a chi-square distribution with $df = 2$,
2. $\frac{2n\bar{W}}{\beta}$ has a chi-square distribution with $df = 2n$, and
3. $\frac{2W_{n+1}}{\beta}$ and $\frac{2n\bar{W}}{\beta}$ are independent.

   The pivot for this model will be:

$$\text{Pivot} = \frac{W_{n+1}}{\bar{W}} = \frac{\left(\frac{2W_{n+1}}{\beta}\right)/2}{\left(\frac{2n\bar{W}}{\beta}\right)/2n}, \quad \text{which is distributed F-distribution with } \ df = 2, 2n$$

The $100(1-\alpha)\%$ P.I. for W_{n+1} is

$$\left(\bar{W}F_{1-\frac{\alpha}{2}}, \bar{W}F_{\frac{\alpha}{2}}\right)$$

which is obtained from

$$P\left[\bar{W}F_{1-\frac{\alpha}{2}} \leq W_{n+1} \leq \bar{W}F_{\frac{\alpha}{2}}\right] = P\left[F_{1-\frac{\alpha}{2}} \leq \frac{W_{n+1}}{\bar{W}} \leq F_{\frac{\alpha}{2}}\right] = 1-\alpha$$

Tolerance Intervals

Confidence intervals are interval estimators of **population parameters** and reflect our uncertainty in estimating these parameters based on a random sample of observations from the population.

Prediction intervals are forecasts or predictions of the value of a **random variable**. The prediction interval reflects our uncertainty in predicting the measured value on a randomly selected unit from the population.

There are many other types of intervals used in production processes and laboratories:

1. **Engineering Tolerance Interval**

   A set of specification limits (S_L, S_U) placed on a product which define an acceptable range of values for the product.

   For example,

   a. A 1.2 Kg box of cereal has $(S_L, S_U) = 1.2 \pm .1$ Kg

   b. A 3 cm diameter piston ring has $(S_L, S_U) = 3 \pm .01$ cm

   c. A 100 ohm resister has $(S_L, S_U) = 100 \pm 5$ ohm

   If the product measurement falls outside of the range (S_L, S_U) then the product is not acceptable for its intended use. The only statistical question is what proportion, p, of the process's output is Acceptable.

   Case 1: Population pdf f is known

$$p = P[Acceptable] = P[Y \epsilon (S_L, S_U)] = \int_{S_L}^{S_U} f(y)dy = p$$

   Case 2: Population pdf f is unknown

   Estimate $p = P[Acceptable]$ using a $100(1-\alpha)$ C.I. based on inspecting n units selected from the process output.

   Let $\hat{p}$ be the proportion of the n units which are acceptable, and the just use $\hat{p}$ to construct a 95% C.I. on p. For example, use the Agresti-Coull C.I. for a proportion.

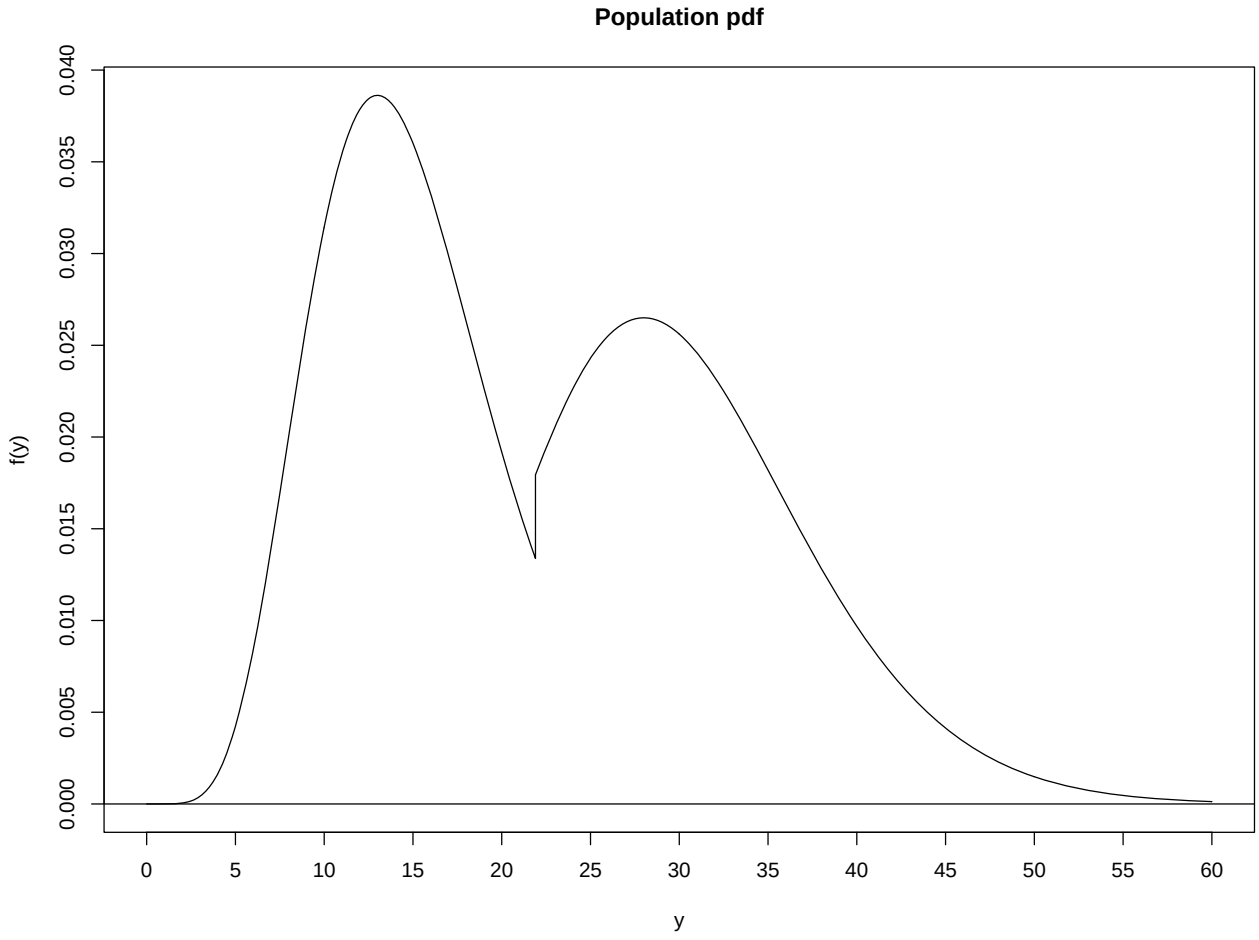
2. **Natural Tolerance of size** $100(1-\alpha)$

When the cdf F for the measurements from a process is completely specified, no unknown parameters, then an interval of values, (T_L, T_U), can we constructed such that

$P[Y\epsilon(T_L, T_U)] = 1 - \alpha$

In fact, one such interval would $T_L = Q(u_1)$ and $T_U = Q(u_2)$, where

$Q(u)$ is the quantile function associated with F and $u_1 + 1 - u_2 = \alpha$.

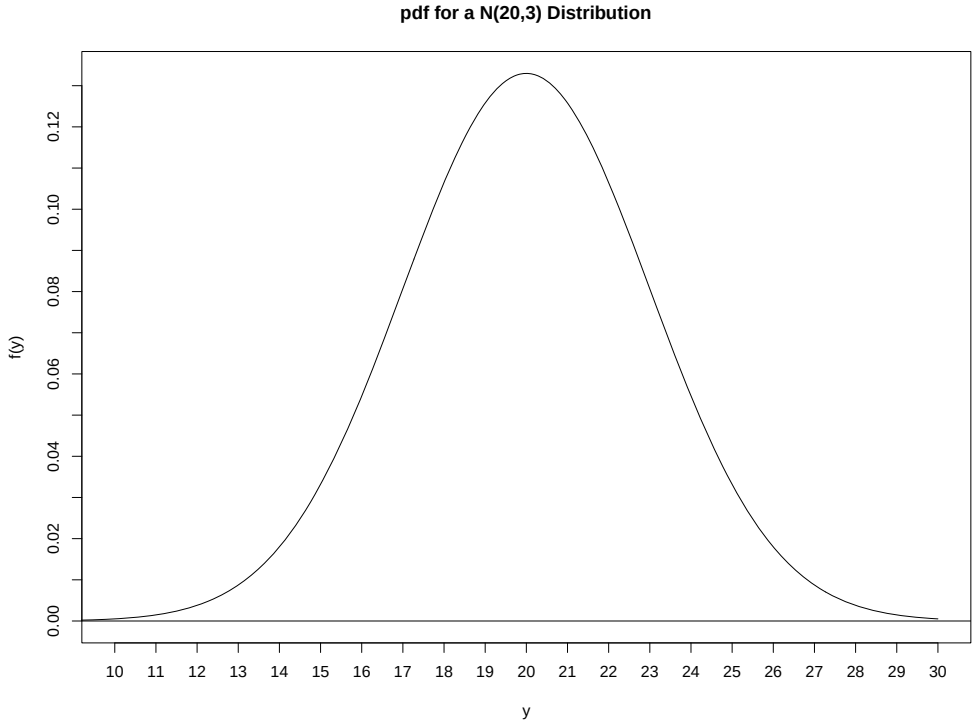


In particular, for a process in which the product characteristic in distributed $N(\mu, \sigma^2)$, with μ and σ known, we have

$Q(u) = \mu + \sigma Z_u$, where Z_u is the $100u$ percentile from a $N(0,1)$ distribution. Thus, the $100(1-\alpha)\%$ natural tolerance would be

$$(T_L, T_U) = \mu \pm \sigma Z_{\frac{\alpha}{2}}, \text{ where we take } u_1 = 1 - u_2 = \frac{\alpha}{2}.$$

For example, with $\mu = 20$ and $\sigma = 3$ we would have a 99% Natural Tolerance Interval would be $20 \pm (3)(Z_{.005}) = 20 \pm (3)(2.58) = (12.26, 27.74)$



3. **Statistical Tolerance Interval** (what you use in practice)

Statistical tolerance intervals are natural tolerance intervals when the population distribution is unknown or contains unknown parameters.

That is, a $100(P,\gamma)\%$ Statistical Tolerance Intervals, T.I., establish limits, $(L_{P,\gamma}, U_{P,\gamma})$ that include $100P\%$ of the responses in a population or from a process with $100\gamma\%$ confidence.

These intervals are used in evaluating production, quality, and service characteristics in many manufacturing and service industries. The T.I. reflects the actual variability of the product or service. T.I.'s are not intervals about a population or process parameter, they are intervals that include a specified portion of the observations from a population or process.

For example, what is the typical systolic blood pressure of health adult females of age 25-50? What would be more useful: a value of mean pressure, μ, or a C.I. on μ or a range of values such that we are 99% confident that 90% of all health adult females of age 25-50 fall into this range?

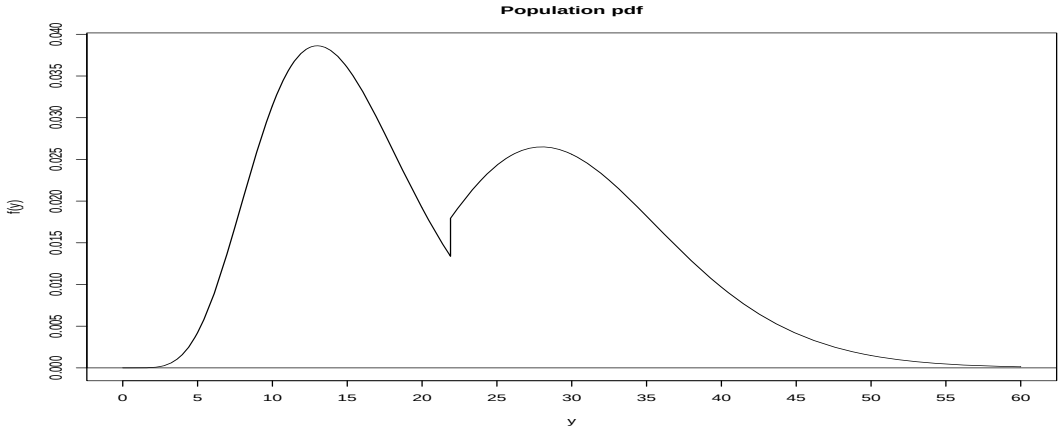
We will consider two examples of parametric T.I.'s and a more general distribution-free T.I.

General expression for a $100(P,\gamma)\%$ T.I. :

A $100(P,\gamma)\%$ T.I $(L_{P,\gamma}, U_{P,\gamma})$ for a population having cdf $F(\cdot)$ and pdf $f(\cdot)$ satifies

$$P\left[A(P,\gamma) \geq P\right] = \gamma \quad \text{where} \quad A(P,\gamma) = \int_{L_{P,\gamma}}^{U_{P,\gamma}} f(y)dy = F(U_{P,\gamma}) - F(L_{P,\gamma}).$$

That is, there is a $100\gamma\%$ probability that at least $100P\%$ of the population falls in the interval $(L_{P,\gamma}, U_{P,\gamma})$.



If f is known completely, no unknown parameters, then just take

$$L_{P,\gamma} = Q(P_1), \quad U_{P,\gamma} = Q(P_2) \quad \text{where} \quad P_2 - P_1 = P$$

Tolerance Intervals for $N(\mu, \sigma^2)$

Case 1: μ and σ Known

If μ and σ are known, then a $\gamma = 1.0$, 100% certainty interval, (L_P, U_P) exists. Namely,

$$A_p = \int_{L_P}^{U_P} \phi(y)dy = \Phi(U_{P,\gamma}) - \Phi(L_{P,\gamma}) = P$$

.

In fact, just let

$$L_P = Q_Y(P_1) = \mu + \sigma Z_{1-P_1}, \quad U_P = Q_Y(P_2) = \mu + \sigma Z_{1-P_2}$$

where $P_2 - P_1 = P$ and

Z_{1-P_1} and Z_{1-P_2} are the upper percentiles from a $N(0,1)$.

If we put equal weight in both tails, that is, $P_1 = 1 - P_2$, then we obtain

$$P_1 = .5(1-P), \quad P_2 = .5(1+P)$$

Case 2: μ and σ Unknown

If μ and σ are unknown, then we must use data to estimate them.

Let $Y_1, \ldots, Y_n$ be $iidN(\mu, \sigma^2)$ r.v.'s.

Then use the following estimators of μ and σ:

$$\hat{\mu} = \bar{Y}, \quad \hat{\sigma} = S$$

Need to find constant $K_{P,\gamma}$ to reflect the uncertainty in using $\hat{\mu}$ and $\hat{\sigma}$ in the intervals:

$$L_{P,\gamma} = \hat{\mu} - K_{P,\gamma}S \quad U_{P,\gamma} = \hat{\mu} + K_{P,\gamma}S$$

where

$$P\left[A(P,\gamma) \geq P\right] = \gamma \text{ and } A(P,\gamma) = \int_{L_{P,\gamma}}^{U_{P,\gamma}} f(y)dy, \text{ with } f(y) \text{ a } N(\mu,\sigma^2) \text{ pdf.}$$

Numerical approximations are used to determine $K_{P,\gamma}$ and a table of values is given on the following page. These tables are from the book, *Statistical Design and Analysis of Experiments*, by Mason, Gunst, and Hess. The values in these tables are slightly different from the values given in the textbook.

One-sided Tolerance Intervals

1-sided tolerance intervals are obtained by just placing all the probability in one tail of the normal distribution. That is,

lower tolerance bound: $L^*_{P,\gamma} = \hat{\mu} - K^*_{P,\gamma}S$

upper tolerance bound: $U^*_{P,\gamma} = \hat{\mu} + K^*_{P,\gamma}S$

Values of $K^*_{P,\gamma}$ for these intervals are also given in the tables on the following pages.

The lower tolerance bound is such that we are $100\gamma\%$ confident that at least $100P\%$ of the population values are greater than $L^*_{P,\gamma}$, yielding $(L^*_{P,\gamma},\ \infty)$

For example, this value could be used as a Warranty. Suppose we are placing on a package of light bulbs the life length of the light bulbs. We are 95% confident that 99% of the company's light bulbs will last at least $L^*_{.99,.95}$ hours: $(L^*_{.99,.95},\ \infty)$

The upper tolerance bound is such that we are $100\gamma\%$ confident that at least $100P\%$ of the population values are less than $U^*_{P,\gamma}$, yielding $(0,\ U^*_{P,\gamma})$

For example, suppose we are producing paint which contains a small amount of lead. We could state that we are 99% confident that 99.5% of our containers of paint will have at most $U^*_{.995,.99}$ units of lead in the container of paint, yielding $(0,\ U^*_{.995,.99})$,

The following R code yields approximations to the exact coefficients.

However, only use when all three of the following conditions hold:

1. $P \geq .95$
2. $\gamma \geq .95$
3. $n > 50$.

```
#Coefficients for One and Two Sided Tolerance Intervals
n =  100
G =  .90
P =  .99
Chi =  qchisq(1-G,n-1)
z =  qnorm((1+P)/2)
K2Side =  sqrt(((n-1)*(n+1)*z^2)/(n*Chi))

za =  qnorm(G)
zb =  qnorm(P)
a =  1-za^2/(2*(n-1))
b =  zb^2-za^2/n
K1Side =  (zb+sqrt(zb^2-a*b))/a
```

Factors for Determining Two-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.900	0.950	0.990	0.900	0.950	0.990	0.900	0.950	0.990
2	15.512	18.221	23.423	31.092	36.519	46.944	155.569	182.720	234.877
3	5.788	6.823	8.819	8.306	9.789	12.647	18.782	22.131	28.586
4	4.157	4.913	6.372	5.368	6.341	8.221	9.416	11.118	14.405
5	3.499	4.142	5.387	4.291	5.077	6.598	6.655	7.870	10.220
10	2.546	3.026	3.958	2.856	3.393	4.437	3.617	4.294	5.610
15	2.285	2.720	3.565	2.492	2.965	3.885	2.967	3.529	4.621
20	2.158	2.570	3.372	2.319	2.760	3.621	2.675	3.184	4.175
25	2.081	2.479	3.254	2.215	2.638	3.462	2.506	2.984	3.915
30	2.029	2.417	3.173	2.145	2.555	3.355	2.394	2.851	3.742
35	1.991	2.371	3.114	2.094	2.495	3.276	2.314	2.756	3.618
40	1.961	2.336	3.069	2.055	2.448	3.216	2.253	2.684	3.524
45	1.938	2.308	3.032	2.024	2.412	3.168	2.205	2.627	3.450
50	1.918	2.285	3.003	1.999	2.382	3.129	2.166	2.580	3.390
60	1.888	2.250	2.956	1.960	2.335	3.068	2.106	2.509	3.297
70	1.866	2.224	2.922	1.931	2.300	3.023	2.062	2.457	3.228
80	1.849	2.203	2.895	1.908	2.274	2.988	2.028	2.416	3.175
90	1.835	2.186	2.873	1.890	2.252	2.959	2.001	2.384	3.133
100	1.823	2.172	2.855	1.875	2.234	2.936	1.978	2.357	3.098
150	1.786	2.128	2.796	1.826	2.176	2.859	1.905	2.271	2.985
200	1.764	2.102	2.763	1.798	2.143	2.816	1.866	2.223	2.921
250	1.750	2.085	2.741	1.780	2.121	2.788	1.839	2.191	2.880
300	1.740	2.073	2.725	1.767	2.106	2.767	1.820	2.169	2.850
350	1.732	2.064	2.713	1.757	2.094	2.752	1.806	2.152	2.828
400	1.726	2.057	2.703	1.749	2.084	2.739	1.794	2.138	2.810
450	1.721	2.051	2.695	1.743	2.077	2.729	1.785	2.127	2.795
500	1.717	2.046	2.689	1.737	2.070	2.721	1.777	2.117	2.783
550	1.713	2.041	2.683	1.733	2.065	2.713	1.770	2.109	2.772
600	1.710	2.038	2.678	1.729	2.060	2.707	1.765	2.103	2.763
650	1.707	2.034	2.674	1.725	2.056	2.702	1.759	2.097	2.755
700	1.705	2.032	2.670	1.722	2.052	2.697	1.755	2.091	2.748
750	1.703	2.029	2.667	1.719	2.049	2.692	1.751	2.086	2.742
800	1.701	2.027	2.664	1.717	2.046	2.688	1.747	2.082	2.736
850	1.699	2.025	2.661	1.715	2.043	2.685	1.744	2.078	2.731
900	1.697	2.023	2.658	1.712	2.040	2.682	1.741	2.075	2.727
950	1.696	2.021	2.656	1.711	2.038	2.679	1.738	2.071	2.722
1000	1.695	2.019	2.654	1.709	2.036	2.676	1.736	2.068	2.718
∞	1.645	1.960	2.576	1.645	1.960	2.576	1.645	1.960	2.576

Factors for Determining One-sided Tolerance Limits

	$\gamma = 0.90$			$\gamma = 0.95$			$\gamma = 0.99$		
	p			p			p		
n	0.90	0.95	0.99	0.90	0.95	0.99	0.90	0.95	0.99
2	10.253	13.090	18.500	20.581	26.260	37.094	103.029	131.426	185.617
3	4.258	5.311	7.340	6.155	7.656	10.553	13.995	17.370	23.896
4	3.188	3.957	5.438	4.162	5.144	7.042	7.380	9.083	12.387
5	2.742	3.400	4.666	3.407	4.203	5.741	5.362	6.578	8.939
10	2.066	2.568	3.532	2.355	2.911	3.981	3.048	3.738	5.074
15	1.867	2.329	3.212	2.068	2.566	3.520	2.521	3.102	4.222
20	1.765	2.208	3.052	1.926	2.396	3.295	2.276	2.808	3.832
25	1.702	2.132	2.952	1.838	2.292	3.158	2.129	2.633	3.601
30	1.657	2.080	2.884	1.777	2.220	3.064	2.030	2.515	3.447
35	1.624	2.041	2.833	1.732	2.167	2.995	1.957	2.430	3.334
40	1.598	2.010	2.793	1.697	2.125	2.941	1.902	2.364	3.249
45	1.577	1.986	2.761	1.669	2.092	2.898	1.857	2.312	3.180
50	1.559	1.965	2.735	1.646	2.065	2.862	1.821	2.269	3.125
60	1.532	1.933	2.694	1.609	2.022	2.807	1.764	2.202	3.038
70	1.511	1.909	2.662	1.581	1.990	2.765	1.722	2.153	2.974
80	1.495	1.890	2.638	1.559	1.964	2.733	1.688	2.114	2.924
90	1.481	1.874	2.618	1.542	1.944	2.706	1.661	2.082	2.883
100	1.470	1.861	2.601	1.527	1.927	2.684	1.639	2.056	2.850
150	1.433	1.818	2.546	1.478	1.870	2.611	1.566	1.971	2.740
200	1.411	1.793	2.514	1.450	1.837	2.570	1.524	1.923	2.679
250	1.397	1.777	2.493	1.431	1.815	2.542	1.496	1.891	2.638
300	1.386	1.765	2.477	1.417	1.800	2.522	1.475	1.868	2.608
350	1.378	1.755	2.466	1.406	1.787	2.506	1.461	1.850	2.585
400	1.372	1.748	2.456	1.398	1.778	2.494	1.448	1.836	2.567
450	1.366	1.742	2.448	1.391	1.770	2.484	1.438	1.824	2.553
500	1.362	1.736	2.442	1.385	1.763	2.475	1.430	1.814	2.540
550	1.358	1.732	2.436	1.380	1.757	2.468	1.422	1.806	2.530
600	1.355	1.728	2.431	1.376	1.752	2.462	1.416	1.799	2.520
650	1.352	1.725	2.427	1.372	1.748	2.456	1.411	1.792	2.512
700	1.349	1.722	2.423	1.368	1.744	2.451	1.406	1.787	2.505
750	1.347	1.719	2.420	1.365	1.741	2.447	1.401	1.782	2.499
800	1.344	1.717	2.417	1.363	1.737	2.443	1.397	1.777	2.493
850	1.343	1.714	2.414	1.360	1.734	2.439	1.394	1.773	2.488
900	1.341	1.712	2.411	1.358	1.732	2.436	1.390	1.769	2.483
950	1.339	1.711	2.409	1.356	1.729	2.433	1.387	1.766	2.479
1000	1.338	1.709	2.407	1.354	1.727	2.430	1.385	1.762	2.475
∞	1.282	1.645	2.326	1.282	1.645	2.326	1.282	1.645	2.326

Lower Tolerance Bound for Exponential Distribution

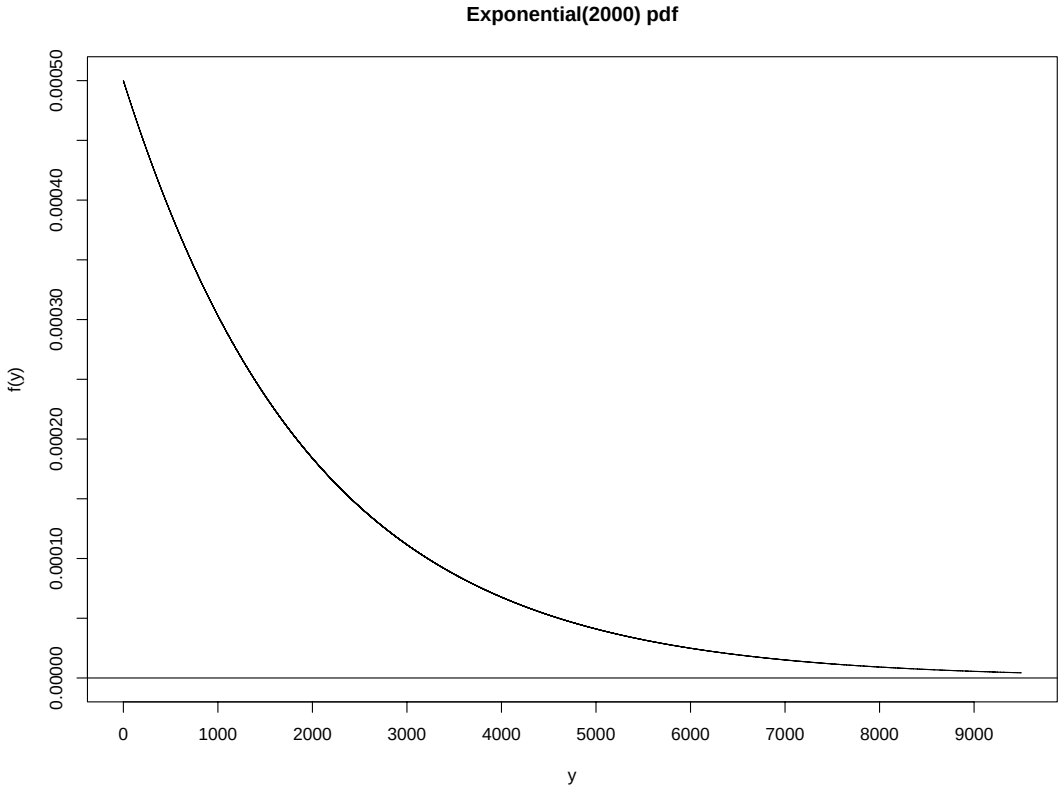
Suppose a company is producing an electronic device which has time to failure modelled by T which has an exponential distribution with average time to failure β. The company wants to set a warranty for the product W such that if the device fails prior to time W they will repair the device without cost to the customer. They want to determine the value for W such that at most 2% of all devices sold will need to be repaired under the warranty. Thus, we need to find W such that $P[T \geq W] \geq .98$.

Case 1: β known

Based on previous repair history, the failure times are known to have an exponential distribution with average time to failure for the device β known with near certainty. Therefore, we have in general the following specification (P=.98 in our example)

$$P[T \geq W_P] = P \Rightarrow \int_{W_P}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt = P \quad \Rightarrow \quad e^{-W_p/\beta} = P \quad \Rightarrow \; W_P = -\beta log(P)$$

If $\beta = 2000$ hours and $P = .98$, we have $W_{.98} = -2000log(.98) = 40.4$ hours.



Case 2: β unknown

The company has made some modifications to the device and thus needs to possibly change the value of W_P. They are certain that the distribution of T is still exponential but the average time to failure for the device β hopefully will have been increased because of improvements in product design.

A survival analysis is conducted on n of the new devices yielding times to failure $T_1, \ldots, T_n$ which are *iid* *Exponential*(β) with β to be estimated from the n data values.

We now need to construct a new lower bound to reflect our uncertainty in the value of β.

Construct $W_{P,\gamma}$ such that we are $100\gamma\%$ confident that $100P\%$ of the population of devices will have times to failure exceeding $W_{P,\gamma}$.

That is, construct $W_{P,\gamma}$ such that with

$$A(W_{P,\gamma}) = \int_{W_{P,\gamma}}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt \quad \text{implies} \quad P\left[A(W_{P,\gamma}) \geq P\right] = \gamma$$

With β known we determined $W_{P,\gamma} = -\beta log(P)$ for all possible values of γ.

To generalize to the case of β unknown we will estimate β using its MLE $\hat{\beta} = \bar{T}$ and

then determine the constant $K_{P,\gamma}$ such that we have

$P\left[A(W_{P,\gamma}) \geq P\right]$, where $W_{P,\gamma} = -\hat{\beta} K_{P,\gamma} log(P)$,

Claim:

$$K_{P,\gamma} = \frac{2n}{\chi^2_{1-\gamma,2n}}$$

where $\chi^2_{1-\gamma,2n}$ is the upper $100(1-\gamma)\%$ percentile from a chi-square distribution with $df = 2n$.

Proof of Claim:

$$A(W_{P,\gamma}) \geq P \iff \int_{W_{P,\gamma}}^{\infty} \frac{1}{\beta} e^{-t/\beta} dt \geq P \iff$$

$$e^{-W_{P,\gamma}/\beta} \geq P \iff -\frac{1}{\beta} W_{P,\gamma} \geq log(P) \iff$$

$$\frac{\hat{\beta}}{\beta} K_{P,\gamma}\ log(P) \geq log(P) \iff \frac{\hat{\beta}}{\beta} K_{P,\gamma} \leq 1 \iff \frac{2n\bar{T}}{\beta} \leq \frac{2n}{K_{P,\gamma}}$$

where $\frac{2n\bar{T}}{\beta}$ is distributed chi-square with df=2n

$$\gamma = P\left[A(W_{P,\gamma}) \geq P\right] \iff \gamma = P\left[\frac{2n\bar{T}}{\beta} \leq \frac{2n}{K_{P,\gamma}}\right] \iff \frac{2n}{K_{P,\gamma}} = \chi^2_{1-\gamma,2n} = qchisq(\gamma, 2n)$$

We thus conclude that a $100(P,\gamma)\%$ Lower Tolerance Bound for an Exponential Distribution is

$$W_{P,\gamma} = -\hat{\beta}\left[\frac{2n}{\chi^2_{1-\gamma,2n}}\right] log(P)$$

END Wednesday 10/27/21

START Friday 10/29/21

Example: Determine a Lower Bound on the time to failure of a device having Exponential failure times such that we are 95% confident that at least 90% of the devices will have failure times greater than the Lower Bound, that is,

Determine a $100(.9, .95)\%$ Lower Tolerance Bound for the device.

Suppose we have $n = 10$ failure times and compute $\hat{\beta} = \bar{T}$.

Then the $100(.9, .95)\%$ Lower Tolerance Bound is given by

$W_{P,\gamma} = -\hat{\beta}\left[\frac{2n}{\chi^2_{1-\gamma,2n}}\right] log(P) = -\bar{T}\left[\frac{20}{\chi^2_{.05,20}}\right] log(.9) = -\bar{T}\left[\frac{20}{31.41043}\right] log(.9) = 0.067\bar{T}.$

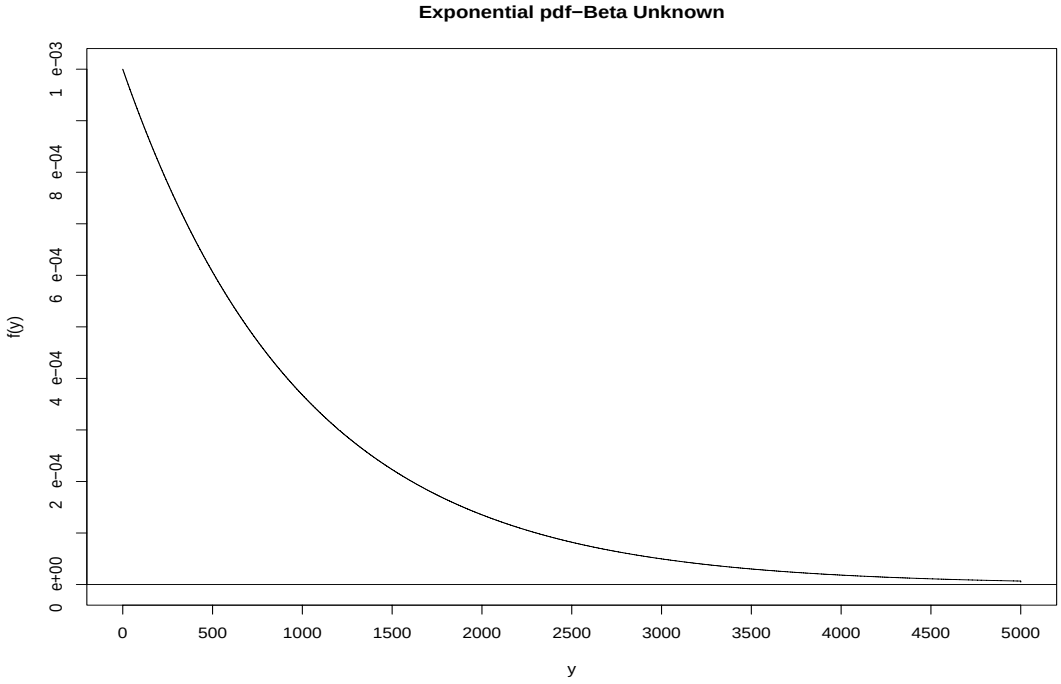
If we would have known β, the Lower Bound would have been

$W_P = -\beta log(P) = -\beta log(.9) = .105\beta.$

Note that for unknown β we have

$E[W_{.9,.95}] = 0.067\beta < .105\beta = W_{.9}.$

Does this relationship seem logical?



Distribution-Free Tolerance Intervals (Bounds) for a Population/Process

In many experiments or studies, the form of the population distribution function F is completely unknown or intractable.

The following procedure yields a distribution-free tolerance interval for a population or process.

Let $Y_1, \ldots, Y_n$ be *iid* with continuous cdf $F(\cdot)$ and pdf $f(\cdot)$.

Let $Y_{(1)} < \ldots < Y_{(n)}$ be the corresponding order statistics for the $Y_i's$.

A $100(P,\gamma)\%$ tolerance interval for the population is given by

$$\left(Y_{(r)},\ Y_{(n-s+1)}\right)$$

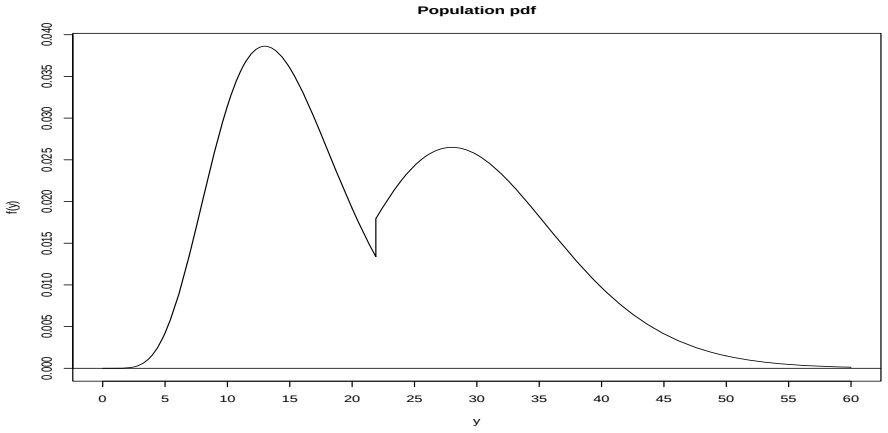
```
   |           |                          |           |
   |--- r ---->|                          |<-- s ----|
---|-----------|------------.......-------|----------|---
 Y(1)        Y(r)                      Y(n-s+1)     Y(n)
```

where r and s are integers satisfying

$$P\left[A(P,\gamma) \geq P\right] = \gamma \quad \text{with}$$

$$A(P,\gamma) = \int_{Y_{(r)}}^{Y_{(n-s+1)}} f(y)dy = F(Y_{(n-s+1)}) - F(Y_{(r)}) = U_{(n-s+1)} - U_{(r)},$$

where $U_{(1)} < \ldots < U_{(n)}$ are the order statistics from n *iid* Uniform on (0,1) r.v.'s.



The tolerance interval is distribution-free by using the probability integral transform theorem, $Y \sim F \Rightarrow F(Y) \sim U_{(0,\ 1)}$ has a uniform on (0, 1) distribution.

The remaining problem is to find the largest integers r and s which yields the narrowest interval $\left(Y_{(r)},\ \ Y_{(n-s+1)}\right)$ such that

$$P\left[Y_{(n-s+1)} - Y_{(r)} \geq P\right] = P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] = \gamma$$

Using properties of the order statistics from an Uniform on (0,1) distribution,

$$P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] = 1 - I_P(n-r-s+1, r+s)$$

where

$$I_P(a,b) = \frac{1}{I_1(a,b)} \int_0^P t^{a-1}(1-t)^{b-1}dt$$

and

$$I_1(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

are incomplete Beta functions.

A few facts about this function will be useful:

1. $I_p(a,b) = 1 - I_{1-p}(b,a)$
2. For $m < n$ integers, $I_p(m, n-m+1) = \sum_{j=m}^{n} \binom{n}{j} p^j (1-p)^{n-j}$
3. $1 - I_p(n-r-s+1, r+s) \geq \gamma$ iff $P[Y \leq m-1] \leq 1-\gamma$

   where Y is distributed $Bin(n, 1-p)$. Thus, we can use the binomial distribution to determine r and s.

- Therefore, $P\left[U_{(n-s+1)} - U_{(r)} \geq P\right] \geq \gamma$ iff $P\left[Y \leq m-1\right] \leq 1-\gamma$

  where $m = r+s$

  Tables for the values of (r,s) are given in Somerville(1958) *Annals of Mathematical Statistics* 29, pp. 599-601. I have included these tables on the next page to obtain $m = r+s$ in the expression: $\left(Y_{(r)},\ \ Y_{(n-s+1)}\right)$.

  To obtain an Upper Tolerance Bound take $r = 0$ yielding $\left(Y_{(0)},\ \ Y_{(n-m+1)}\right)$

  to obtain an Lower Tolerance Bound take $s = 0$ yielding $\left(Y_{(m)},\ \ Y_{(n+1)}\right)$

  where $Y_{(0)}$ is the **population minimum** and $Y_{(n+1)}$ is the **population maximum.**

  When population values are $(0,\ \ \infty)$,

  Upper Tolerance Bound is $\left(0,\ \ Y_{(n-m+1)}\right)$ and lower tolerance bound is $\left(Y_{(m)},\ \ \infty\right)$

Example Based on a random sample of $n = 130$ data values: $Y_1, \ldots, Y_{130}$,

find an interval of values

$[L_{(P,\gamma)}, U_{(P,\gamma)}]$

such that we are 95% certain that at least 90% of the population values are in this interval,

$[L_{(P,\gamma)}, U_{P,\gamma}]$.

That is, find a $100(.9, .95)\%$ tolerance interval for the population.

From the table with $P = .90$, $\gamma = .95$, and $n = 130$, we have $r + s = m = 8$.

Thus, one choice would be to take $r = 4$ and $s = 4$.

Using this choice, the $100(.9, .95)\%$ tolerance interval would be

$[Y_{(4)}, Y_{(130-4+1)}] = [Y_{(4)}, Y_{(127)}]$.

That is, we are 95% certain that at least 90% of the population values are between

$$Y_{(4)} \text{ and } Y_{(127)}.$$

The following R code will compute the value of m given in the tables on the next page along with the actual value of γ:

```
n= 130
G= .95
P= .90
m= 0
imin= 0
i= 0
ans= 0
anst= 0
r= 1:n
ans= pbinom(r-1,n,1-P)
while(i<n)
{
i= i+1
if(ans[i]<=1-G)  anst[i]= ans[i]
if(ans[i]>1-G) anst[i]= -1
}
ansmax= max(anst)
imax= which(anst==ansmax)
m= imax
coverage= 1-ans[m]
out = cbind(m, coverage)
out
# m coverage
  8 .9544028
```

Annals of Mathematical Statistics Vol 29 (1958)

Tolerance Intervals (P, γ): $[X_{(r)}, X_{(n-s+1)}]$

Values of $m = r + s$ such that we may assert with confidence at least γ that 100 P percent of a population lies between the rth smallest and the sth largest of a random sample of n from that population (continuous distribution function assumed)

n	P																								
	$\gamma = 0.50$					$\gamma = 0.75$					$\gamma = 0.90$					$\gamma = 0.95$					$\gamma = 0.99$				
	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99	.50	.75	.90	.95	.99
50	25	12	5	2	0	22	10	3	1	—	20	9	2	1	—	19	8	2	—	—	16	6	1	—	—
55	28	14	5	3	0	25	12	4	2	—	23	10	3	1	—	21	9	2	—	—	19	7	1	—	—
60	30	15	6	3	0	27	13	4	2	—	25	11	3	1	—	24	10	2	1	—	21	8	1	—	—
65	33	16	6	3	0	30	14	5	2	—	27	12	4	1	—	26	11	3	1	—	23	9	2	—	—
70	35	17	7	3	1	32	15	5	2	—	30	13	4	1	—	28	12	3	1	—	25	10	2	—	—
75	38	19	7	4	1	35	16	6	2	—	32	14	4	1	—	30	13	3	1	—	27	10	2	—	—
80	40	20	8	4	1	37	17	6	3	—	34	15	5	2	—	33	14	4	1	—	30	11	2	—	—
85	43	21	8	4	1	39	19	7	3	—	37	16	5	2	—	35	15	4	1	—	32	12	3	—	—
90	45	22	9	4	1	42	20	7	3	—	39	17	5	2	—	37	16	5	1	—	34	13	3	1	—
95	48	24	9	5	1	44	21	7	3	—	41	18	6	2	—	39	17	5	2	—	36	14	3	1	—
100	50	25	10	5	1	47	22	8	3	—	44	20	6	2	—	42	18	5	2	—	38	15	4	1	—
110	55	27	11	5	1	51	24	9	4	—	48	22	7	3	—	46	20	6	2	—	43	17	4	1	—
120	60	30	12	6	1	56	27	10	4	—	53	24	8	3	—	51	22	7	2	—	47	19	5	1	—
130	65	32	13	6	1	61	29	11	5	—	58	26	9	3	—	56	25	8	3	—	52	21	6	2	—
140	70	35	14	7	1	66	31	12	5	1	62	28	10	4	—	60	27	8	3	—	56	23	6	2	—
150	75	37	15	7	1	71	34	12	6	1	67	31	10	4	—	65	29	9	3	—	61	26	7	2	—
170	85	42	17	8	2	81	39	14	7	1	77	35	12	5	—	74	33	11	4	—	70	30	9	3	—
200	100	50	20	10	2	95	46	17	8	1	91	42	15	6	—	88	40	13	5	—	84	36	11	4	—
300	150	75	30	15	3	144	70	26	12	2	139	65	23	10	1	136	63	22	9	1	130	58	19	7	—
400	200	100	40	20	4	193	94	35	17	3	187	89	32	15	2	184	86	30	13	1	177	80	27	11	—
500	250	125	50	25	5	242	118	45	22	3	236	113	41	19	2	232	109	39	17	2	224	103	35	14	1
600	300	150	60	30	6	292	143	55	26	4	284	136	51	23	3	280	133	48	21	2	272	126	44	18	1
700	350	175	70	35	7	341	167	65	31	5	333	160	60	28	4	328	156	57	26	3	319	149	52	22	2
800	400	200	80	40	8	390	192	74	36	6	382	184	69	32	5	377	180	66	30	4	367	172	61	26	2
900	450	225	90	45	9	440	216	84	41	7	431	208	79	37	5	425	204	75	35	4	415	195	70	30	3
1000	500	250	100	50	10	489	241	94	45	8	480	233	88	41	6	474	228	85	39	5	463	219	79	35	3

Tolerance Interval (P, γ): $[X_{(1)}, X_{(n)}]$

Confidence γ with which we may assert that 100 P percent of the population lies between the largest and smallest of a random sample of n from that population (continuous distribution assumed)

n	$P = .50$	$P = .75$	$P = .90$	$P = .95$	$P = .99$	n	$P = .75$	$P = .90$	$P = .95$	$P = .99$
3	.50	.16	.03	.01	.00	17	.95	.52	.21	.01
4	.69	.26	.05	.01	.00	18	.96	.55	.22	.01
5	.81	.37	.08	.02	.00	19	.97	.58	.25	.02
6	.89	.47	.11	.03	.00	20	.98	.61	.26	.02
7	.94	.56	.15	.04	.00	25	.99	.73	.36	.03
8	.96	.63	.19	.06	.00	30	1.00−	.82	.45	.04
9	.98	.70	.23	.07	.00	40		.92	.60	.06
10	.99	.76	.26	.09	.00	50		.97	.73	.09
11	.99	.80	.30	.10	.01	60		.99	.81	.12
12	1.00−	.84	.34	.12	.01	70		.99	.87	.16
13		.87	.38	.14	.01	80		1.00−	.91	.19
14		.90	.42	.15	.01	90			.94	.23
15		.92	.45	.17	.01	100			.96	.26
		.94	.49	.19	.01					

Comparison of Distribution-Free to Parametric Tolerance Intervals

Because there is a Distribution-Free method to obtain Tolerance Intervals, why bother with using the specific Tolerance Intervals for given families of distributions?

The major reason is that the Distribution-Free Tolerance Intervals may be considerably wider than the corresponding interval for a particular family because the Distribution-Free Tolerance Intervals must be applicable to a wide-variety of distributions.

Suppose the population or process cdf F is from a $N(\mu, \sigma^2)$ family of distributions.

Normal based Tolerance Interval: $\bar{Y} \pm K_{p,\gamma} S$

Distribution-free Tolerance Interval: $\left(Y_{(r)},\ Y_{(n-s+1)}\right)$

For $100(.9, .95)\%$ Tolerance Intervals, compare the widths of the normal-based tolerance interval to the distribution-free tolerance interval. Because the widths are random variables we will consider the expected widths for comparison under the assumption that the population distribution is $N(\mu, \sigma^2)$.

$W_N = 2K_{P,\gamma} S$ $\qquad$ $W_{DF} = Y_{(n-s+1)} - Y_{(r)}$ which yields

$E[W_N] = 2K_{P,\gamma} E[S] = 2K_{P,\gamma} \frac{\sigma\sqrt{2}}{\sqrt{n-1}} \frac{\Gamma(n/2)}{\Gamma((n-1)/2)}$

$E[W_{DF}] = E[Y_{(n-s+1)}] - E[Y_{(r)}]$

with $E[Y_{(r)}] = \mu + \sigma \frac{n!}{(n-r)!(r-1)!} \int_{-\infty}^{\infty} y(.5 - \Phi(y))^{r-1}(.5 + \Phi(y))^{n-r}\phi(y)dy,$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the $N(0,1)$ cdf and pdf, respectively.

Evaluating these expressions yields the following results:

n	$E[W_{DF}]$	$E[W_N]$	$E[W_{DF}]/E[W_N]$
50	4.498σ	3.978σ	1.15
100	4.094σ	3.741σ	1.10
200	3.753σ	3.591σ	1.05

Thus, even for a relatively large value of n, $n = 100$ the Distribution-Free Tolerance Interval is 10% wider than the normal-based Tolerance Interval.

Alternative Approaches to Constructing C.I.'s, P.I.'s, and T.I.'s

Two other approaches to constructing C.I.'s, P.I.'s, and T.I.'s are to transform the data to normality and then use a normal-based approach or use bootstrap procedures on the untransformed data.

Box-Cox Transformations to Normality

Suppose the population distributions is non-normal, that is, Y has cdf $F_Y(\cdot)$ which is non-normal but the Box-Cox transformation $X = g(Y)$ results in X having a cdf $F_X(\cdot)$ which is approximately a normal distribution. How can Intervals generated for X yield a corresponding interval for Y?

1. **Tolerance Interval for Distribution of Y based on $X = g(Y)$**

   Because X is approximately normally distributed we obtain a $100(P,\gamma)\%$ T.I. for the distribution for X:

   $(L^*_{P,\gamma}, U^*_{P,\gamma}) \Rightarrow \quad \gamma = P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P].$

   Case 1: Increasing Function: Ex $\theta = 2$

   Let $g^{-1}(\cdot)$ be the inverse of g where g is a strictly increasing function, a $100(P,\gamma)\%$ T.I. for the distribution of Y is

   $(L_{P,\gamma}, U_{P,\gamma}) = \left(g^{-1}(L^*_{P,\gamma}), g^{-1}(U^*_{P,\gamma})\right).$

   This result follows from:

   $F_Y(y) = P[Y \leq y] = P[g^{-1}(X) \leq y] = P[g(g^{-1}(X)) \leq g(y)] = P[X \leq g(y)] = F_X(g(y)) \Rightarrow$

   $$\begin{aligned} P[F_Y(U_{P,\gamma}) - F_Y(L_{P,\gamma}) \geq P] &= P[F_Y(g^{-1}(U^*_{P,\gamma})) - F_Y(g^{-1}(L^*_{P,\gamma})) \geq P] \\ &= P[F_X(g(g^{-1}(U^*_{P,\gamma}))) - F_X(g(g^{-1}(L^*_{P,\gamma}))) \geq P] \\ &= P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P] \\ &= \gamma \end{aligned}$$

   Case 2: Decreasing Function: Ex: $\theta = -0.5$

   Let g be a strictly decreasing function, a $100(P,\gamma)\%$ T.I. for the distribution of Y is

   $(L_{P,\gamma}, U_{P,\gamma}) = \left(g^{-1}(U^*_{P,\gamma}), g^{-1}(L^*_{P,\gamma})\right).$

   This result follows from:

   $F_Y(y) = P[Y \leq y] = P[g^{-1}(X) \leq y] = P[g(g^{-1}(X)) \geq g(y)] = P[X \geq g(y)] = 1 - F_X(g(y))$

   $$\begin{aligned} P[F_Y(U_{p,\gamma}) - F_Y(L_{p,\gamma}) \geq P] &= P[F_Y(g^{-1}(L^*_{p,\gamma})) - F_Y(g^{-1}(U^*_{p,\gamma})) \geq P] \\ &= P[1 - F_X(g(g^{-1}(L^*_{P,\gamma}))) - 1 + F_X(g(g^{-1}(U^*_{P,\gamma}))) \geq P] \\ &= P[F_X(U^*_{P,\gamma}) - F_X(L^*_{P,\gamma}) \geq P] \\ &= \gamma \end{aligned}$$

2. **Prediction Interval for Distribution of Y based on $X = g(Y)$**

The same results from above would apply to prediction intervals.

Suppose $Y_1, \ldots, Y_n$ are *iid* but have a non-normal distribution but $g(Y_i) = X_i$ is approximately normally distributed.

Case 1: g a strictly increasing function

Let $\bar{X}$ and S_X be the sample mean and standard deviation of the $X_i s$

Let $\bar{X} \pm t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}$ be a $100(1-\alpha)\%$ P.I. for X_{n+1}, that is,

$$\begin{aligned} 1-\alpha &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq X_{n+1} \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq g(Y_{n+1}) \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right) \leq Y_{n+1} \leq g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right)\right] \Rightarrow \end{aligned}$$

$$\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right), g^{-1}\left(X + t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right)\right)$$

is a $100(1-\alpha)\%$ P.I. for Y_{n+1}

Case 2: g a strictly decreasing function

$$\begin{aligned} 1-\alpha &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq X_{n+1} \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}} \leq g(Y_{n+1}) \leq \bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right] \\ &= P\left[g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right) \geq Y_{n+1} \geq g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\frac{n+1}{n}}\right)\right] \Rightarrow \end{aligned}$$

$$\left(g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right), g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S_X \sqrt{\tfrac{n+1}{n}}\right)\right)$$

is a $100(1-\alpha)\%$ P.I. for Y_{n+1}

3. Confidence Intervals for Population μ and σ Using Transformations

The results from above **DO NOT** in general apply to confidence intervals for μ, σ, and most other parameters related to population moments. However, transformations are valid for Quantiles, such as the population median, $Q(.5)$.

Suppose $Y_1, \ldots, Y_n$ are *iid* but have a non-normal distribution but $g(Y_i) = X_i$ is approximately normally distributed with g a strictly increasing function. We want to obtain a $100(1-\alpha)$ C.I. for μ_Y.

Let $\bar{X} \pm t_{\frac{\alpha}{2}} S/\sqrt{n}$ be a $100(1-\alpha)\%$ C.I. for μ_X.

Is $\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right), g^{-1}\left(X + t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right)\right)$ a $100(1-\alpha)\%$ C.I. for μ_Y?

We cannot conclude this because, in general, $\mu_X \neq g(\mu_Y)$ and hence $g^{-1}(\mu_X) \neq \mu_Y$.

This results from the following:

$\mu_X = E[X] = E[g(Y)] \neq g(E[Y]) = g(\mu_Y)$, thus, $g^{-1}(\mu_X) \neq \mu_Y$.

Using a Taylor series expansion of $g(y)$ about μ_Y we obtain

$X = g(Y) = g(\mu_Y) + g'(\mu_Y)(Y - \mu_Y) + \frac{1}{2} g''(\mu_Y)(Y - \mu_Y)^2 + R \Rightarrow$

$$\begin{aligned} \mu_X = E[X] = E[g(Y)] &= g(\mu_Y) + g'(\mu_Y)E[(Y - \mu_Y)] + \frac{1}{2} g''(\mu_Y)E[(Y - \mu_Y)^2] + E[R] \\ &= g(\mu_Y) + \frac{1}{2} g''(\mu_Y)Var(Y) + E[R] \\ &\approx g(\mu_Y) \end{aligned}$$

provided $Var(Y)$ is small and/or $g''(\mu_Y)$ is small, and $E[R]$ is small.

In many cases, neither $Var(Y)$ is small nor $E[R]$ is small.

This leads to a very bad approximation of μ_X using $g(\mu_Y)$ and hence the C.I. for μ_Y obtained from $\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right), g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}\ \right)\right)$ will not be an appropriate C.I. for μ_Y.

An alternative approach in such situations is to attempt to find a C.I. for μ_Y directly from the distribution of Y.

If this is not possible then the **bootstrap C.I.** may be an alternative methodology for obtaining the confidence interval.

Example Let Y have a lognormal distribution, that is, $Y = e^W$, where W is $N(\mu, \sigma^2)$.

Let

$$X = g(Y) = log(Y)$$

Then, X has a $N(\mu, \sigma^2)$ distribution with

$\mu_X = E[X] = E[log(Y)] = \mu$ and

$\mu_Y = E[Y] = e^{\mu + \frac{1}{2}\sigma^2}$.

How close is μ_X to $g(\mu_Y) = log(\mu_Y)$?

$log(\mu_Y) = \mu + \frac{1}{2}\sigma^2 = \mu_X + \frac{1}{2}\sigma^2 \quad \Rightarrow \quad \mu_X = log(\mu_Y) - \frac{1}{2}\sigma^2$.

Therefore, if σ^2 is large relative to μ_X, then

μ_X will not be very close to $g(\mu_Y) = log(\mu_Y)$.

A $100(1-\alpha)\%$ C.I. for μ_X is $\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n} \;\; , \;\; \bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n} \;\right)$

Inverting the end points of the C.I. for μ_X, we obtain:

$$\left(g^{-1}\left(\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}\;\right), \;\; g^{-1}\left(\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}\;\right)\right) = \left(e^{\bar{X} - t_{\frac{\alpha}{2}} S/\sqrt{n}}, \;\; e^{\bar{X} + t_{\frac{\alpha}{2}} S/\sqrt{n}}\right)$$

This interval would be a C.I. for the parameter

$$e^{\mu_X} = e^{log(\mu_Y) - \frac{1}{2}\sigma^2} = \mu_Y e^{-\frac{1}{2}\sigma^2}$$

and not a C.I. for μ_Y.

To obtain an approximate C.I. for μ_Y would involve simultaneous C.I.'s for μ and σ^2

4. **Confidence Intervals for Population Quantiles: $Q(u)$ Using Transformations**

Suppose we want a C.I. for $Q_Y(u)$ for the distribution of the r.v. Y but standard techniques are not feasible.

However, there exists a transformation, of Y, $X = g(Y)$ such that a $100(1-\alpha)\%$ C.I. can be constructed for $Q_X(u)$

C.I. on $Q_X(u) = (L_x,\ U_x)$

A. If g is an increasing function then recall that $Q_X(u) = g(Q_Y(u))$ thus a $100(1-\alpha)\%$ C.I. on $Q_Y(u)$ is given by

$(g^{-1}(L_x),\ g^{-1}(U_x))$

B. If g is a decreasing function then recall that $Q_X(u) = g(Q_Y(1-u))$ thus a $100(1-\alpha)$ C.I. on $Q_Y(u)$ is given by

$(g^{-1}(U_x^*),\ g^{-1}(L_x^*))$, where

$(L_x^*,\ U_x^*)$ is a $100(1-\alpha)\%$ on $Q_X(1-u)$

Bootstrap Confidence Intervals for Parameters Related to cdf

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 $\hat{\theta}_n = \hat{\theta}(X_1, \ldots, X_n)$.

Furthermore, we want to construct a $100(1-\alpha)\%$ C.I. for θ based on the pivot $R_n = \hat{\theta}_n - \theta$.

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

We need to determine the sampling distribution of the pivot $R_n = \hat{\theta}_n - \theta$ in order to be able to determine its percentiles, R_α, which are necessary in order to obtain a $100(1-\alpha)\%$ C.I. for θ.

We cannot use exact mathematical derivations to derive the true sampling distribution of $\hat{\theta}_n - \theta$ because the cdf $F(\cdot)$ is unknown. Also, the asymptotic distribution may not provide an adequate approximation of the true sampling distribution of $\hat{\theta}_n - \theta$ when n is small to moderate in size.

An alternative to these two approaches is the **bootstrap procedure** which will provide an approximation to the sampling distribution of the pivot, $\hat{\theta}_n - \theta$, in those situations where we can write θ as a function of the cdf, that is, $\theta = g(F(\cdot))$.

To obtain the sample estimator, we simply replace the true cdf $F(\cdot)$ with the empirical (sample) cdf $\hat{F}(x)$ in $\theta = g(F(\cdot))$ to obtain $\hat{\theta}_n = g(\hat{F}(\cdot))$.

We want to obtain the sampling distribution of $\hat{\theta}_n - \theta$ using simulated data from the edf $\hat{F}(\cdot)$ in place of the true cdf $F(\cdot)$.

Let $\hat{\theta}_D$ be the value of $\hat{\theta}_n$ computed from the edf, $\hat{F}$, that is, from the data, $X_1,\ X_2,\ \ldots,\ X_n$.

The sampling distribution of $\hat{\theta}_n - \theta$, will be approximated by the sampling distribution of $\hat{\theta}^*_n - \hat{\theta}_D$, where $\hat{\theta}^*_n$ is the value of $\hat{\theta}_n$ from a bootstrap sample.

Thus, we have replaced $\hat{\theta}_n$ with $\hat{\theta}^*_n$, its value from the bootstrap sample and θ with its estimator from the edf, $\hat{F}$, $\hat{\theta}_D$.

More formally, if $G(\cdot)$ is the sampling cdf of $\hat{\theta}_n - \theta$, that is,

$G(y) = P[\hat{\theta}_n - \theta \leq y]$,

then the bootstrap simulation estimator of $G(y)$ based on B bootstrap samples is given by

$$\widehat{G}_B(y) = \frac{1}{B}\sum_{i=1}^{B} I[\hat{\theta}^*_i - \hat{\theta}_D \leq y]$$

the proportion of the B samples in which $\hat{\theta}^*_i - \hat{\theta}_D$ are less than or equal to y and

where $\hat{\theta}^*_1, \hat{\theta}^*_2, \ldots, \hat{\theta}^*_B$

are B estimates of θ based on the B bootstrap samples of size n taken from the original data set.

STOP Friday 10/29/21

The approximation of $\widehat{G}_B(u)$ to $G(u)$ contains two sources of error:

1. the difference between $\widehat{G}(u)$ and $G(u)$ due to data variability from the original population
2. the difference between $\widehat{G}_B(u)$ and $\widehat{G}(u)$ due to a finite number of bootstrap samples.

We will approximate the percentiles of the sampling distribution of the pivot $R_n = \hat{\theta}_n - \theta$ using the sample percentiles from the

B values of $R_n^* = \hat{\theta}^* - \hat{\theta}_D$: $R_{n,1}^*, R_{n,2}^*, \cdots, R_{n,B}^*$

This results from the fact that if $X_1, \ldots, X_M$ are iid with cdf $K(\cdot)$ and if $X_{(i)}$ denotes the ith ordered value of the $X_i's$, then

$E[X_{(i)}] \approx K^{-1}(\frac{i}{M+1})$.

Thus, a sensible estimator of $Q(p) = K^{-1}(p)$ is $X_{((M+1)p)}$, provided $(M+1)p$ is an integer.

So we will estimate the $100pth$ quantile of the pivot $R_n = \hat{\theta}_n - \theta$ by

the $(B+1)pth$ ordered value in $R_{(1)}^*, R_{(2)}^*, \cdots, R_{(B)}^*$,

that is, $R_{((B+1)p)}^* = \hat{\theta}_{((B+1)p)}^* - \hat{\theta}_D$,

where $\hat{\theta}_{((B+1)p)}^*$ is the $(B+1)pth$ ordered value of $\hat{\theta}_{(1)}^*, \hat{\theta}_{(2)}^*, \cdots, \hat{\theta}_{(B)}^*$,.

In our development we will select B such that $(B+1)p$ is an integer.

For example, suppose $n = 20$ and we want to obtain

the $100(.025) = 2.5$th percentile and $100(.975) = 97.5$th percentile.

With B=9999 then we would use

the $(B+1)(.025) = 250$th ordered value of R_n^*, $R_{(250)}^*$ to estimate the 2.5th percentile and

$(B+1)(.975) = 9750$th ordered value of R_n^*, $R_{(9750)}^*$ to estimate the 97.5th percentile

START Monday 7/11/21

Basic Bootstrap Confidence Limits

For a given confidence level $1-\alpha$, and using the Pivot $= \hat{\theta}_n - \theta$,

replace w/ sample estimate i.e. MLE.

we need to find values $L_{\frac{\alpha}{2}}$ and $U_{\frac{\alpha}{2}}$ such that

replace w/ bootstrap estimate

$P(L_{\frac{\alpha}{2}} \leq \hat{\theta}_n - \theta \leq U_{\frac{\alpha}{2}}) \approx 1-\alpha$ which would yield

$\left(\hat{\theta}_n - U_{\frac{\alpha}{2}},\ \hat{\theta}_n - L_{\frac{\alpha}{2}}\right)$ as the $100(1-\alpha)$ C.I. for θ.

Because the sampling distribution of the Pivot $= \hat{\theta}_n - \theta$ is unknown, we will convert the problem to the following with

$\hat{\theta}_n^*$ in place of $\hat{\theta}_n$ and → $\hat{\theta}_n^*$ is $\hat{\theta}_n$ estimate from bootstrap

$\hat{\theta}_D$ in place of θ in the Pivot and use the bootstrap samples to obtain the necessary percentiles. $\hat{\theta}_n$ MLE $\hat{\theta}_D$ estimate

Let $R^* = \hat{\theta^*}_{ni} - \hat{\theta}_D$.

Generate B bootstrap values for $R^*:\ R_1^*,\ R_2^*,\ \ldots, R_B^*$ and then order these values:

$$R_{(1)}^* \leq R_{(2)}^* \leq \ldots \leq R_{(B)}^*$$

For a given confidence level $1-\alpha$, find values $L_{\frac{\alpha}{2}}^*$ and $U_{\frac{\alpha}{2}}^*$ such that

$P(L_{\frac{\alpha}{2}}^* \leq \hat{\theta}_n^* - \hat{\theta}_D \leq U_{\frac{\alpha}{2}}^*) \approx 1-\alpha \Rightarrow$ (using the results on the previous page)

$L_{\frac{\alpha}{2}}^* = R_{((B+1)(\frac{\alpha}{2}))}^* = \theta_{((B+1)(\frac{\alpha}{2}))}^* - \hat{\theta}_D$ and

$U_{\frac{\alpha}{2}}^* = R_{((B+1)(1-\frac{\alpha}{2}))}^* = \hat{\theta}_{((B+1)(1-\frac{\alpha}{2}))}^* - \hat{\theta}_D$.

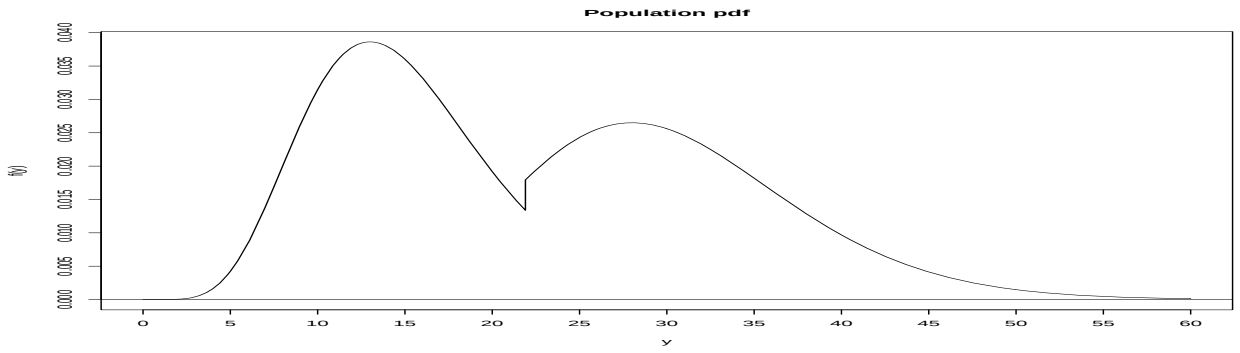
Therefore, an approximate $100(1-\alpha)\%$ C.I. for θ is given by

$$\left(\hat{\theta}_D - R_{((B+1)(1-\frac{\alpha}{2}))}^*,\quad \hat{\theta}_D - R_{((B+1)(\frac{\alpha}{2}))}^*\right) = \left(2\hat{\theta}_D - \hat{\theta}_{((B+1)(1-\frac{\alpha}{2}))}^*,\quad 2\hat{\theta}_D - \hat{\theta}_{((B+1)(\frac{\alpha}{2}))}^*\right)$$

The accuracy of the approximation depends on the size of B, thus one typically takes $B \geq 10000$.

Accuracy also depends upon the extent to which the distribution of $\hat{\theta}_n^* - \hat{\theta}_D$ agrees with that of $\hat{\theta}_n - \theta$.

If the distribution of $\hat{\theta}_n - \theta$ does depend on unknown parameters, then alternative formulations of the C.I.'s must be invoked. The reference *Bootstrap Methods and their Applications*, Davison and Hinkley, discuss this situation in detail, including diagnostics to determine if the pivot depends on θ. We will consider this situation following the next example.



Example The following example is from *Computer Intensive Statistical Methods*, by Hjorth.

In a large installation of electric bulbs, all the bulbs are planned to be replaced regularly after 1200 hours. In order to form an opinion about this strategy, the probability that a bulb will survive 1200 hours is of interest. Let T be the time to failure of this type of bulb. The parameter of interest is

$\theta = P[T \geq 1200] = 1 - F(1200) \Rightarrow \theta = g(F) = 1 - F(1200)$, where $F(\cdot)$ is the cdf of T.

A limited test of the bulbs is conducted and the following 20 life times are observed:

1354	1552	1766	1325	2183	1354	1299	627	695	2586
2420	71	2195	1825	159	1577	3725	884	1014	965

From the data, we compute $\hat{\theta}_n = 1 - \hat{F}(1200) = \frac{13}{20} = .65 = \theta_D$

because 13 of the 20 bulbs failed after 1200 hours.

We will compute an approximate 95% C.I. for θ using the following bootstrap program: **boot1,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)

n= length(x)

m= sum(x>1200)

thest = m/n

B = 9999

thestS = numeric(B)

thestS = rep(0,times =B)

for (i in 1:B)

thestS[i] = sum(sample(x,replace=T)>1200)/20

RS = sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

The approximate 95% C.I. for θ is (thL, thU) = (.45, .85).

Suppose we also want to compute an approximate 95% confidence interval for the median time to failure:
$\theta = Median$.

We would proceed similarly as above
using the following bootstrap program: **boot2,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)

n= length(x)

thest = median(x)

B = 9999

thestS = numeric(B)

thestS = rep(0,times =B)

for (i in 1:B)

thestS[i] = median(sample(x,replace=T))

RS= sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

The point estimate is $\hat{\theta}_n = 1354$.

The approximate 95% C.I. for θ is

$$(\hat{\theta}_L,\ \hat{\theta}_U) = (thL,\ thU) = (912.5,\ 1718.5)$$

Using our nonparametric C.I. for the median, an approximated 95% c.i. for the median is

$$\left(\hat{\theta}_L,\ \hat{\theta}_U\right) \ = \ \left(Y_{(k)},\ Y_{(n-k+1)}\right) = \left(Y_{(6)},\ Y_{(15)}\right) = (965,\ 1825)$$

The two intervals are reasonable close considering that we only have $n = 20$ data values.

Studentized Bootstrap C.I. (Improvement from above when we have additional information.)

If the distribution of $\hat{\theta}_n - \theta$ depends on unknown parameters, then the basic bootstrap procedure may not be very accurate.

For example, suppose $\theta = Q(u)$ for some value of u, $0 < u < 1$.

$\hat{\theta}_n = \hat{Q}(u)$ and $\sqrt{n}\left(\hat{\theta}_n - \theta\right)$

has asymptotic standard error:

$$SE_{Asy}(\hat{\theta}) = \frac{\sqrt{u(1-u)}}{f(Q(u))} = \frac{\sqrt{u(1-u)}}{f(\theta)}$$

which is a function of θ

A method which regains some of the lost accuracy is to use the *Studentized* version of $\hat{\theta}_n - \theta$:

$$Z = \frac{\hat{\theta}_n - \theta}{\sqrt{\widehat{V}}}$$; Pivot in studentized bootstrap

where $\widehat{V}$ is an estimate of $Var(\hat{\theta}_n)$.

The **Studentized Bootstrap C.I. for** θ is given by

$$\left(\hat{\theta}_D - \sqrt{\widehat{V}_D}\ Z^*_{((B+1)(1-\frac{\alpha}{2})},\ \ \hat{\theta}_D - \sqrt{\widehat{V}_D}\ Z^*_{((B+1)(\frac{\alpha}{2})}\ \right)$$

where $\hat{\theta}_D$ is the sample estimate of θ and $\widehat{V}_D$ is the sample estimate of V.

$Z^*_{((B+1)(1-\frac{\alpha}{2})}$ and $Z^*_{((B+1)(\frac{\alpha}{2})}$

are the $(B+1)(\frac{\alpha}{2})$ and $(B+1)(1-\frac{\alpha}{2})$ ordered values obtained from

B bootstrap samples of Z: $Z^*_1, \ldots, Z^*_B$.

The difficulty is that we need to obtain a form for $\hat{V}$.

In some cases, the form will be known or we can use the asymptotic variance for $\hat{V}$.

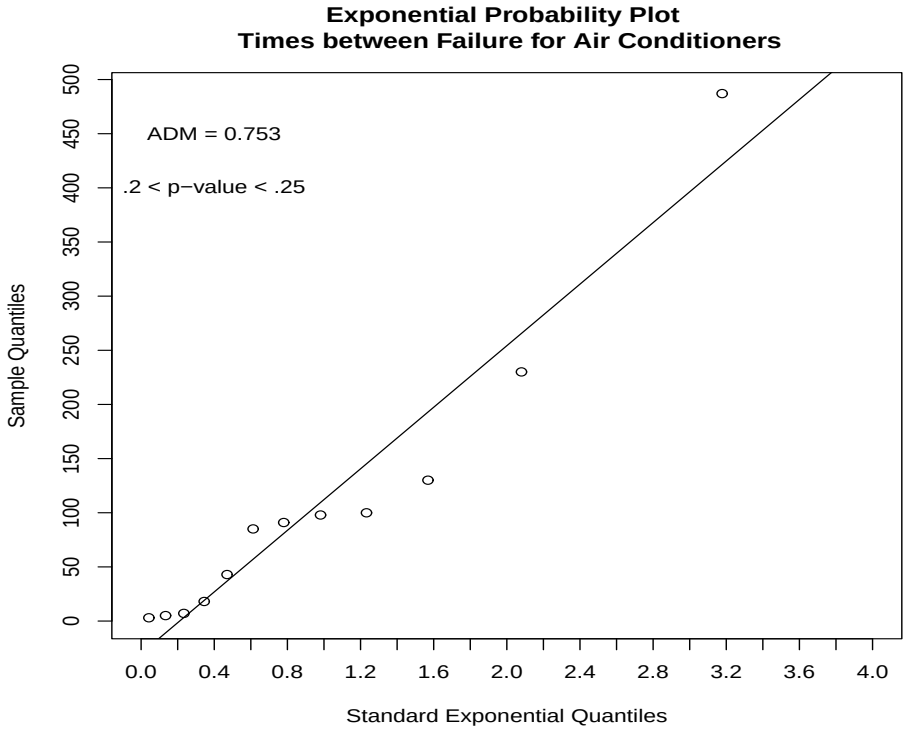
Davison and Hinkley provide a general form for $\hat{V}$ involving influence functions and the nonparametric delta method.

Example A study yielded the n=12 times between failure of air-conditioning units, for which we wish to estimate the average time between failures:

3	5	7	18	43	85
91	98	100	130	230	487

Suppose we model the times to failure as an exponential distribution (see Exponential Reference Distribution Plot) and compute

AD=0.753 with $.20 < p - value < .25$.



For the exponential model, the unknown parameter $\beta = E[T]$ and $Var[T] = \beta^2 \ \Rightarrow \ SE(\hat{\beta}) = \frac{\beta}{\sqrt{n}}$.

If we did have prior information that $T_1,\ T_2, \ldots,\ T_n$ were iid Exp(β) r.v.s, and used $\bar{T} - \mu$ as the pivot for the C.I. for μ, then the sampling distribution of $\bar{T} - \mu$ would depend on the unknown parameter because $\text{Var}(\bar{T} - \mu) = \beta/\sqrt{n}$.

Using the knowledge that the data is from an exponential distribution, the MLE's are $\hat{\beta} = \bar{T}$ and $\hat{V} = \hat{\beta}^2/n$.

Thus, let $\hat{V} = \bar{T}^2/n$ which yields

$$Z = \frac{\hat{\theta}_D - \theta}{\sqrt{\hat{V}}} = \frac{\bar{T} - \beta}{\bar{T}/\sqrt{n}} = \frac{\sqrt{n}(\bar{T} - \beta)}{\bar{T}}$$

An exact C.I.'s for β can be obtained using the fact that the sampling distribution of $2n\bar{T}/\beta$ has a chi-squared distribution with df=2n.

To illustrate the studentized bootstrap technique, the following R code was used **bootexp,ci.R:**

bootexp,ci.R:

```
x= c(3,5,7,18,43,85,91,98,100,130,230,487)

n= length(x)

thest = mean(x)

V = thest**2/n

B = 9999

W = numeric(B)

W = rep(0,times =B)

for (i in 1:B)

W[i] = mean(sample(x,replace=T))

Z = sqrt(n)*(W-thest)/W

Z = sort(Z)

LZ = Z[250]

UZ = Z[9750]

thL = thest-UZ*sqrt(V)

thU = thest-LZ*sqrt(V)
```

We obtain $Z^*_{((B+1)(1-\frac{\alpha}{2})} = Z^*_{(9750)} = 1.494$ and $Z^*_{((B+1)(\frac{\alpha}{2}))} = Z^*_{(250)} = -4.474$.

Thus, an approximate 95% C.I. for β is

$$\left(\hat{\theta}_n - \sqrt{\hat{V}} Z^*_{((B+1)(1-\frac{\alpha}{2})}, \quad \hat{\theta} - \sqrt{\hat{V}} Z^*_{((B+1)(\frac{\alpha}{2})} \right) =$$

$$\left(108.08 - \sqrt{973.5006}(1.494), \quad 108.08 - \sqrt{973.5006}(-4.474) \right) = (61.5, 247.7)$$

The C.I. for β using the fact we are sampling from an exponential distribution and using the pivot $2n\bar{T}/\beta$ is given by

$$\left(\frac{2n\bar{T}}{\chi^2_{.025}}, \frac{2n\bar{T}}{\chi^2_{.975}} \right) = \left(\frac{24(108.08)}{39.364}, \frac{24(108.08)}{12.401} \right) = (65.9, 209.2)$$

The bootstrap C.I. and Exact C.I. are surprising close considering that the C.I.'s are based on only $n = 12$ data values.

— Basic Bootstrap example -> NOTE: Not Studentized

Example In the light bulb example on page 61, we wanted to estimate the value of bootstrap b/c

$\theta = P[T \geq 1200] = g(F) = 1 - F(1200)$, where $F(\cdot)$ is the cdf of T.

If we take $F(t) = 1 - e^{-t/\beta}$, exponential model, then the parameter of interest is we don't know the

$\theta = 1 - F(1200) = e^{-1200/\beta}$, with $\hat{\beta} = \bar{T} = 1478.8$, standard error of $\hat{\theta}$.

we obtain $\hat{\theta}_n = e^{-1200/1478.8} = 0.444$

To obtain a basic bootstrap C.I. for θ we will use the following program **boot3,ci.R**:

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,2420,71,2195,1825,159,1577,3725,884,1014,965)
n = length(x)

mn= mean(x)

thest = exp(-1200/mn)

R = 9999
thestS = numeric(R)
thestS = rep(0,times =R)

for (i in 1:R)
thestS[i] = exp(-1200/mean(sample(x,replace=T)))

RS = sort(thestS-thest)

LRS = RS[250]

URS = RS[9750]

thL = thest-URS

thU = thest-LRS
```

From the R output we have:

$R^*_{(1000)(.025)} = RS[250] = -.100, \quad R^*_{(1000)(.975)} = RS[9750] = .080, \quad \hat{\theta}_D = .444$

Then, our approximate 95% C.I. for θ is

$$\begin{aligned} \left(\hat{\theta}_D - R^*_{.975}, \;\; \hat{\theta}_D - R^*_{.025}\right) &= (.444 - (.080), \;\; .444 - (-.100)) \\ &= (.364, \;\; .544) \end{aligned}$$

Example of Studentized Bootstrap C.I. for θ

Assuming that an exponential model for the light bulb data is appropriate, a studentized bootstrap C.I. for the parameter

$\theta = P[T \geq 1200] = e^{-1200/\beta}$

could be calculated provided we are able to obtain an approximation to the variance of $\hat{\theta} = e^{-1200/\hat{\beta}}$.

If the standard deviation of $\bar{T}$ is small, then a 1-term Taylor expansion of $g(y) = e^{-1200/y}$ about μ_T would be appropriate. This yields

$$\hat{\theta} = g(\bar{T}) \approx g(\mu_T) + g^{'}(\mu_T)(\bar{T} - \mu_T) = e^{-1200/\mu_T} + \left(\frac{1200}{\mu_T^2}\right) e^{-1200/\mu_T} (\bar{T} - \mu_T)$$

From which we obtain

$$V = Var(\hat{\theta}) \approx \left(\frac{1200}{\mu_T^2}\right)^2 e^{-2400/\mu_T} Var\left(\bar{T} - \mu_T\right) \Rightarrow \hat{V} = \left(\frac{1200}{\bar{T}^2}\right)^2 e^{-\frac{2400}{\bar{T}}} \frac{S^2}{n}$$

where $\hat{\mu}_T = \bar{T}$ and S^2 is the sample variance of the $T_i's$.

We can now bootstrap $Z = (\hat{\theta}_n - \hat{\theta}_D)/\sqrt{V}$.

From each bootstrap sample, compute $\bar{T}^*, S^*, V^*, Z^*$ and

then obtain the upper and lower sample $\frac{\alpha}{2}$ percentiles of Z^*.

Using the following program **boot4,ci.s**, we can obtain a 95% C.I. for θ

```
x= c(1354,1552,1766,1325,2183,1354,1299,627,695,2586,
      2420,71,2195,1825,159,1577,3725,884,1014,965)
n = length(x)
mn= mean(x)
thhat = exp(-1200/mn)

S2 = var(x)

Vest= ((1200/(mn**2))**2)*(exp(-2400/mn))*(S2/n)

R = 9999
z = numeric(R)
z = rep(0,times =R)
for (i in 1:R)
{t= sample(x,replace=T)

V= ((1200/(mean(t)**2))**2)*(exp(-2400/mean(t)))*(var(t)/n)

z[i] = (exp(-1200/mean(t))-thhat)/sqrt(V)}

z = sort(z)

L = z[250]

U = z[9750]

thL = thhat-U*sqrt(Vest)

thU = thhat-L*sqrt(Vest)
```

From the output, we have

$\bar{T} = 1478.8; \quad \hat{V}_D = .002246486; \quad \hat{\theta} = 1 - \hat{F}(1200) = e^{-1200/\bar{T}} = e^{-1200/1478.8} = .444$

$Z^*_{((R+1)(1-\frac{\alpha}{2})} = Z^*_{250} = -2.030352$

$Z^*_{((R+1)(\frac{\alpha}{2})} = Z^*_{9750} = 2.193155$

Then, our approximate 95% C.I. for θ is

$$\left(\hat{\theta}_n - \sqrt{\widehat{V}_D} Z^*_{((R+1)(1-\frac{\alpha}{2})}, \quad \hat{\theta} - \sqrt{\widehat{V}_D} Z^*_{((R+1)(\frac{\alpha}{2})} \quad \right) =$$

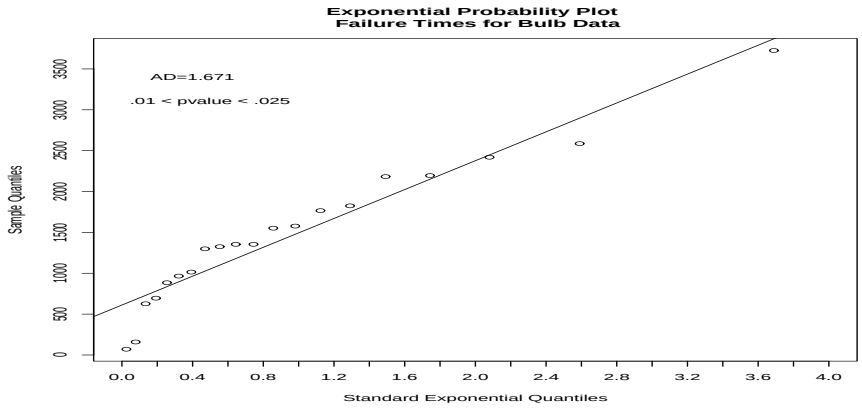
$$\left(.444 - (\sqrt{.002246486})(2.193155), \quad .444 - (\sqrt{.002246486})(-2.030352)\right) = (.340, \quad .540)$$

We have calculated 5 different confidence intervals for the parameter $\theta = P[T > 1200]$:

Method	Confidence Interval
Clapper-Pearson(with θ a population proportion)	(.408, .846)
Wilson C.I.(with θ a population proportion)	(.433, .819)
Basic Bootstrap (with θ a population proportion)	(.450, .850)
Basic Bootstrap (with $\theta = 1 - e^{-1200/\beta}$)	(.361, .547)
Studentized Bootstrap (with $\theta = 1 - e^{-1200/\beta}$)	(.340, .540)

Why do you think there is such a difference between the intervals obtained by treating the parameter θ as a proportion in comparison to the intervals obtained by treating the parameter θ as a function of β?

For the exponential distribution, the two representation of θ are identical, why are the two C.I.'s for θ so different?



Parametric Bootstrap

In some cases a parametric model may be known for the data but the sampling distribution of some statistics is of such a complex nature that the sampling distribution cannot be determined mathematically and/or the sample size n is too small to invoke asymptotic properties of the sampling distribution. In these types of situations, a **parametric bootstrap** procedure can be implemented.

Example Suppose we are studying the effectiveness of a new insecticide for controlling the damage on various crops due to infestation of fire ants. The treatments are applied to one acre plots of land planted with the crops of interest. The response variable is the amount of useful crop harvested from the field.

The parameter of interest is the coefficient of variation: $\theta = \frac{\sigma}{\mu}$ since the crops are very different in terms of the mean yield.

From many previous studies it is known that the crop yields follow a log-normal distribution but the parameters, (μ, σ), will be unknown.

The maximum likelihood estimator of θ is denoted as $\hat{\theta} = \frac{\hat{\sigma}}{\hat{\mu}}$.

The sampling distribution of $\hat{\theta}$ is known asymptotically but there is no explicit expression for the sampling distribution for small sample sizes. A C.I. for θ can be estimated using the parametric bootstrap.

In our previous discussion the bootstrap samples were obtained by random sampling with replacement from the n original data values, that is, sampling from a population having cdf $\widehat{F}$, the edf obtained from the original data.

In the case of the parametric bootstrap, the unknown parameters in the cdf are estimated using the MLEs computed from the original data set. Samples are then simulated from the parametric cdf with the unknown parameters replaced with their MLEs. We thus run the simulation B times yielding

B bootstrap samples $(Y_1^*, Y_2^*, \ldots, Y_n^*)_i \quad \text{for} \quad i = 1, 2, \ldots, B.$

The modification from the procedures used previously to determine a bootstrap C.I. for θ is that when a parametric model was unknown we were generating the B bootstrap samples of n observations from the original data $Y_1, \ldots, Y_n$. In the parametric bootstrap we are generating the B bootstrap samples from a log-normal distribution with the values of the parameters μ and σ determined by the MLEs computed from the original data.

START Wednes 11/3/21

Example of Parametric Bootstrap

Suppose we have data from a logistic distribution with (location, scale) parameters $(\beta,\ \gamma)$ unknown. A $100(1-\alpha)\%$ is desired on the coefficient of variation, $\theta = \frac{\sigma}{\mu} = \frac{\pi\gamma/\sqrt{3}}{\beta}$.

The sampling distribution of the MLE for θ is unknown for small to moderate sized samples. In particular, n=25 would be too small to apply the asymptotic results for the functions of MLE's.

Let $X_1, X_2, \ldots, X_{25}$ be the data from which we compute the MLE's $\hat{\beta}$ and $\hat{\gamma}$.

Next we use R to generate $B = 10,000$ samples of size $n = 25$ from a logistic distribution with parameters $\hat{\beta}$ and $\hat{\gamma}$.

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_1^*, X_2^*, \ldots, X_{25}^*)_1\ \Rightarrow\ (\hat{\beta}_1^*,\ \hat{\gamma}_1^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_1^*/\sqrt{3}}{\hat{\beta}_1^*}$$

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_2^*, X_2^*, \ldots, X_{25}^*)_2\ \Rightarrow\ (\hat{\beta}_2^*,\ \hat{\gamma}_2^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_2^*/\sqrt{3}}{\hat{\beta}_2^*}$$

$$\vdots$$

$$\text{rlogis}(25,\hat{\beta},\hat{\gamma}) \Rightarrow\ (X_1^*, X_2^*, \ldots, X_{25}^*)_B\ \Rightarrow\ (\hat{\beta}_B^*,\ \hat{\gamma}_B^*)\ \Rightarrow\ \hat{\theta}_1^* = \frac{\pi\hat{\gamma}_B^*/\sqrt{3}}{\hat{\beta}_B^*}$$

From the B estimates of θ: $\hat{\theta}_1,\ \hat{\theta}_2,\ \ldots, \hat{\theta}_B$ we could then construct the C.I. on θ using a bootstrap confidence interval.

Suppose we have 25 observations from a logistic distribution with (Location, Scale) = $(\beta,\ \gamma)$ unknown

```
Data:  16.20  9.37 25.80  9.55 12.86 15.34 18.08 10.76 14.92  9.75 17.10 13.97
       15.08  9.24 11.99 13.60  8.16 12.82 12.89 13.59 16.23 14.19  9.03  9.58  13.68
```

Using the following Rcode: **parambootlogistic,ci.R**, we will obtain a 95% c.i. for the coefficient of determination, CV = $\frac{\sigma}{\mu}$

Because CV is a scale type parameter, an appropriate pivot would be $R = \frac{\widehat{CV}}{CV}$.

First find $R_{\alpha/2}$ and $R_{1-\alpha/2}$ such that $P[R_{\alpha/2} \le R \le R_{1-\alpha/2}] \approx 1-\alpha$.

Obtain estimates by simulating B bootstrap samples of size n and computing

$R_1^*,\ R_2^*,\ \ldots,\ R_B^*$, where $R_i^* = \frac{\widehat{CV}_i^*}{\widehat{CV}_D}$ and

$\widehat{CV}_D$ is the value of CV computed from the original data.

The approximate $100(1-\alpha)$ C.I. for CV would be

$$\left(\frac{\widehat{CV}_D}{R_{((B+1)(1-\alpha/2))}},\ \frac{\widehat{CV}_D}{R_{((B+1)(\alpha/2))}}\right)$$

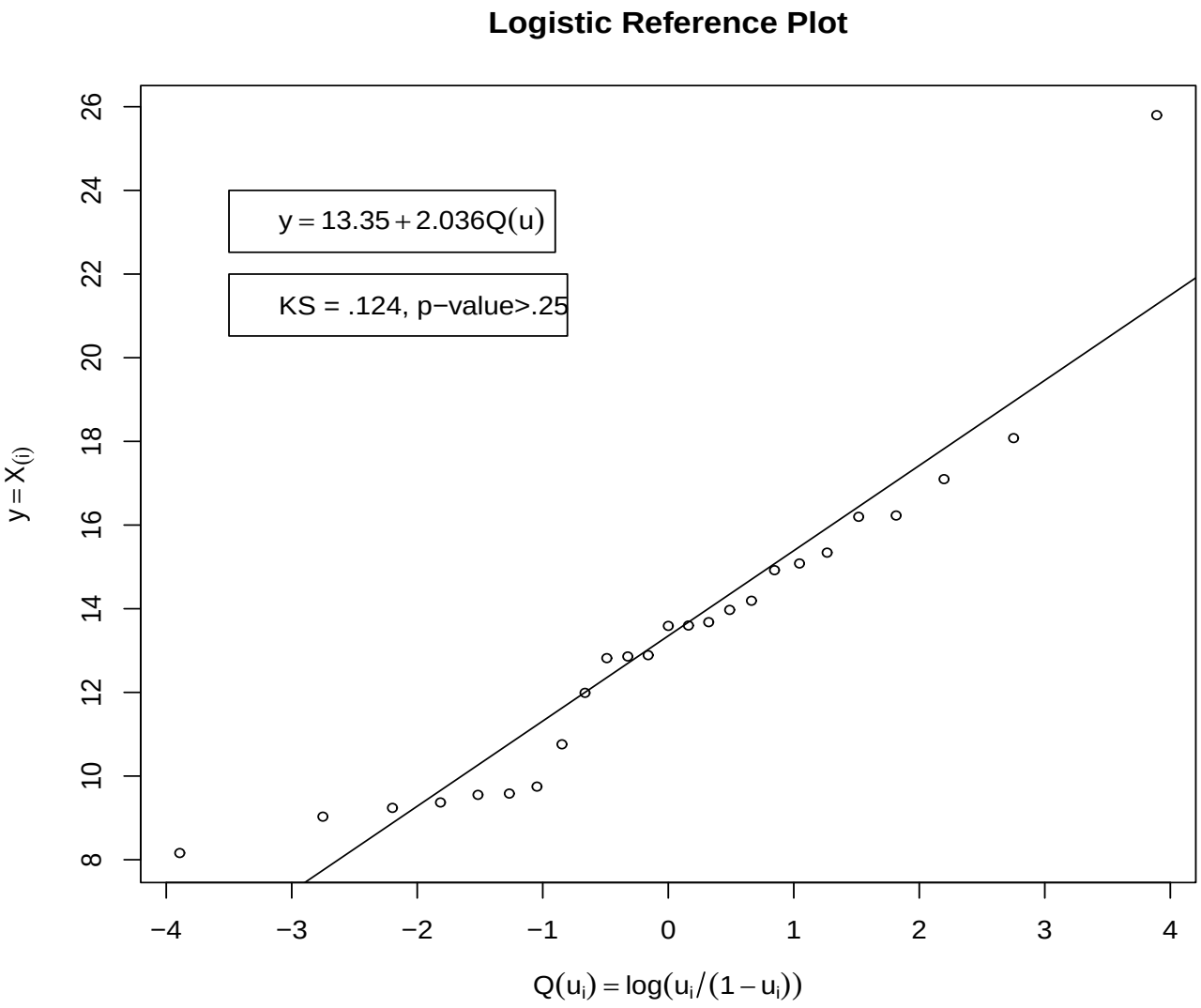
The logistic distribution has location and scale parameters $(\beta,\ \gamma)$ and quantile function:

$$Q(u) = \beta + \gamma log(u/(1-u))$$

thus the quantile function for the standard member is $(\beta = 0,\ \gamma = 1)$: $Q(u) = log(u/(1-u))$

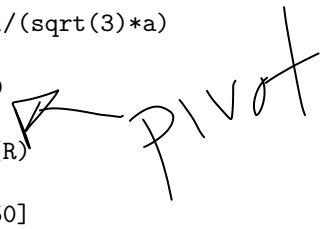
The Logistic Reference Plot for the 25 data values is given below along with the approximate K-S test yielding a p-value from the R-code

```
library(MASS)
mleestD = coef(fitdistr(x,"logistic"))
aD = mleestD[1]
bD = mleestD[2]
ks.test(x,"plogis",aD,bD)
D = 0.12448, p-value = 0.7888
```



parambootlogistic,ci.R

```
library(MASS)

x = c(16.20,9.37,25.80,9.55,12.86,15.34,18.08,10.76,14.92,9.75,17.10,13.97,
15.08,9.24,11.99,13.60,8.16,12.82,12.89,13.59,16.23,14.19,9.03,9.58,13.68)

#obtain MLE of the a=location and b=scale parameters in logistic model

mleestD = coef(fitdistr(x,"logistic"))
aD = mleestD[1]
bD = mleestD[2]
cvD = bD*pi/(sqrt(3)*aD)

n =  length(x)
B  =  9999
W  =  matrix(0,B,n)
cv =  numeric(B)
cv =  rep(0,B)
a =  numeric(B)
a =  rep(0,B)
b =  numeric(B)
b =  rep(0,B)
mleest =  matrix(0,B,2)

{
for (i in 1:B)
W[i,]  =  rlogis(n,aD,bD)
}

{
for (i in 1:B)
mleest[i,] = coef(fitdistr(W[i,],"logistic"))
}

a = mleest[,1]

b = mleest[,2]

cv =  b*pi/(sqrt(3)*a)

R = cv/cvD

R =  sort(R)

L  =  R[250]

U  =  R[9750]

ci = c(cvD/U, cvD/L)

ci
```



The point estimator for the CV is 0.2742 and a 95% C.I. for the CV is (0.202, 0.408)

Skip rest of HD. End @ 12:27- cn 11/5 Lecture

Problems Encountered in Using Bootstrap Procedures

Great care must be taken in using bootstrap procedures, both nonparametric and parametric bootstrap methods. Many inappropriate applications of bootstrap procedures are conducted due to the attitude that there are no assumptions necessary in order to apply a bootstrap procedure. This is far from the truth.

An excellent discussion of some of these problems can be found in Hinkley and Davison's book, *Bootstrap Methods and Their Applications*

A quote from their book, "The error in resampling methods is generally a combination of statistical error and simulation error."

The basic bootstrap idea is to approximate a quantity $c(F)$, where F is the population/process cdf, by the estimate $c(\widehat{F})$, where $\widehat{F}$ is either a parametric or nonparametric estimate of F based on the observed data, $Y_1, \ldots, Y_n$.

The statistical error in the bootstrap procedure is the difference between $c(\widehat{F})$ and $c(F)$. The problem that may arise is that the sampling distribution of the elected pivot may depend on unknown parameters. They describe procedures for checking empirically if this problem exists and remedies when it does. The more basic consideration is that the observed data does not adequately represent the population either distributionally or with respect to outliers.

The simulation errors can generally removed by selecting a large value for B. The only question is the size of B to guarantee a specified degree of accuracy. They discuss the issue of how to select the value of B in a number of situations.

Various conditions are given under which the bootstrap procedure will produce a mathematically justifiable estimate of the sampling distribution of the statistic of interest.

The following example is provided to demonstrate that bootstrap procedures are not universally applicable.

Example Suppose a random sample $Y_1, \ldots, Y_n$ is selected from a uniform on $(0, \theta)$ distribution. The MLE of θ is $\hat{\theta} = Y_{(n)}$.

The pivot for obtaining a C.I. on θ is $PV_n = \frac{n(\theta - \hat{\theta})}{\theta}$.

It is easily shown that the asymptotic distribution of PV_n is a standard exponential distribution, $F(y) = 1 - e^{-y}$

This would suggest that in the bootstrap procedure the pivot should be computed as $PV_n^* = \frac{n(\hat{\theta}_D - \hat{\theta}^*)}{\hat{\theta}_D}$, where $\hat{\theta}_D$ is $\hat{\theta}$ computed from the original data and $\hat{\theta}^*$ is $\hat{\theta}$ computed from each bootstrap sample.

We can easily determine: $P[PV_n^* = 0|\widehat{F}] = P[\hat{\theta}^* = \hat{\theta}_D|\widehat{F}] = 1 - (1 - \frac{1}{n})^n$

Thus, $lim_{n\to\infty} Pr[PV_n^* = 0|\widehat{F}] = 1 - e^{-1}$

Therefore, the limiting distribution PV_n^* cannot be a standard exponential distribution.

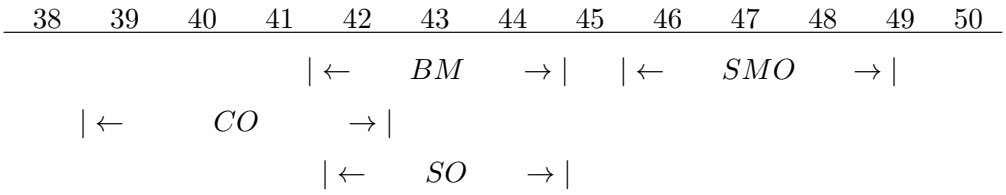
The above illustrates care must be taken in making sure the required regularity conditions prior to using a bootstrap procedure.

Simultaneous Confidence Intervals

In many studies or experiments, there will be more than one population/process of interest to the researcher. Consider the following example. A large percentage of the dietary energy in the bodies of infants is provided by lipids. Lipids are a class of hydrocarbon-containing organic compounds. The following data on total polyunsaturated fats(%) was reported for infants who were randomized to four different feeding regimens: BM (breast milk), CO corn-oil-based formula, SO (soy-oil-based formula), or SMO (soy-marine-oil based formula).

Regimen	n	$\bar{X}$	S	95% C.I. on μ_i
BM	18	43.0	3.5	(41.27, 44.73)
CO	13	40.4	3.3	(38.42, 42.38)
SO	17	43.1	3.2	(41.46, 44.74)
SMO	14	47.1	3.2	(45.27, 48.93)

The researcher determined that there was significant evidence that the data from the four regimen was from a normal distribution. He then proceeded to construct 95% confidence intervals on the mean percentage of polyunsaturated fat in the infants for each of the four regimens, as reported in the above table.



Based on the above confidence intervals, the researcher concluded that there was significant evidence at the 95% level that SMO produced a mean percentage of polyunsaturated fat in the infants that was significantly higher than the other regimens.

A statistician reviewed the results and told the researcher that he may be incorrect in his conclusions. She stated that the above confidence intervals are **individual** confidence intervals and that the researcher needed **simultaneous** confidence intervals in order to reach his stated conclusions.

The researcher was attempting to make a statement about all k (k=4, in the example) parameters: $\theta_1, \theta_2, \ldots, \theta_k$ simultaneously not as separate estimates of the individual parameters. That is, the above intervals were derived using an inference procedure based on

$$\text{Individual } 100(1-\alpha)\% \text{ C.I.'s:} \qquad P\left[\theta_i \epsilon \left(L_i, U_i\right)\right] = 1-\alpha \qquad \text{for all } \; i = 1, \ldots, k$$

In order to reach the conclusion stated by the researcher it is necessary to construct confidence intervals based on the following probability statement:

$$\text{Simultaneous } 100(1-\alpha)\% \text{ C.I.'s:} \qquad P\left[\theta_i \epsilon \left(L_i, U_i\right) \quad \text{for all } \; i = 1, \ldots, k\right] = 1-\alpha$$

Suppose $100(1-\alpha)\%$ individual C.I.s are constructed on k parameters: $\theta_1, \ldots, \theta_k$.

What is the simultaneous coverage probability, γ, of these k intervals? That is, compute

$$\gamma = P\left[\theta_i \epsilon\, (L_i, U_i) \quad \text{for} \quad i = 1, \ldots, k\right]$$

Let A_i be the event $\{\theta_i \epsilon\, (L_i, U_i)\}$

If the intervals (L_i, U_i) are independent random variables for $i = 1, \ldots, k$ with coverage probability $1-\alpha_i$ then

$$\gamma = P\left[A_1 \cap A_2 \cap \cdots \cap A_k\right] = \prod_{i=1}^{k} P\left[A_i\right] = \prod_{i=1}^{k} (1-\alpha_i) = (1-\alpha)^k \quad \text{if} \quad \alpha_i \equiv \alpha$$

The following table will demonstrate that the simultaneous coverage probability can be considerably smaller than the individual coverage probabilities.

k	1	2	3	4	5	$\cdots$	13	14
$1-\alpha$	0.95	0.95	0.95	0.95	0.95	$\cdots$	0.95	0.95
γ	0.95	0.903	0.857	0.815	0.774	$\cdots$	0.513	0.488

After examining the above table, the researcher then asked the statistician how he could construct simultaneous C.I.s for the parameters. The statistician suggested the following procedure but emphasized that it would only be **valid if the individual C.I.s are independent.**

Determining α to Obtain Specified Value for γ

To obtain a $100\gamma\%$ simultaneous C.I. for k parameters using a procedure which generates $100(1-\alpha)\%$ independent individual C.I.s for k parameters, just set the value of $\alpha/2$ at

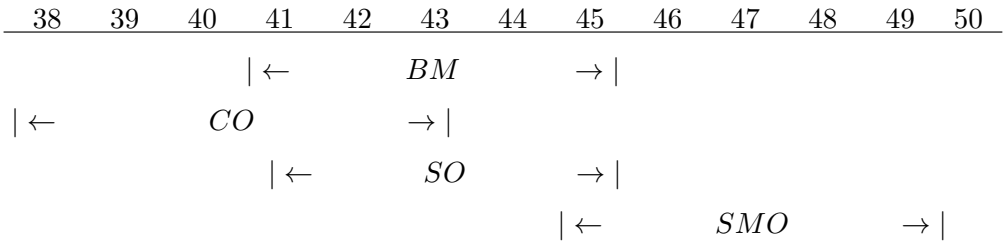
$$1-\alpha = \gamma^{1/k} \;\Rightarrow\; \alpha/2 = .5\left(1-\gamma^{1/k}\right)$$

k	1	2	3	4	5	$\cdots$	13	14
γ	0.9500	0.9500	0.9500	0.9500	0.9500	$\cdots$	0.9500	0.9500
$1-\alpha$	0.9500	0.9747	0.9830	0.9873	0.9898	$\cdots$	0.9960	0.9963
$\alpha/2$	.0250	0.01266	0.008476	0.006371	0.005103	$\cdots$	0.001969	0.001829

Example:

In the lipid example, the researcher wanted a 95% simultaneous C.I. on the $k = 4$ means. Therefore, he would need to use $t_{.00635}$ in the C.I. construction: $\bar{X} \pm t_{\frac{\alpha}{2}} S/\sqrt{n}$ in place of $t_{.025}$, as was done in constructing 4 individual C.I.s. The impact of this replacement can be seen in the following table:

Regimen	n	$\bar{X}$	S	95% Indiv. C.I. on μ_i	95% Simul. C.I. on $\mu_i s$	$t_{.025}$	$t_{.00635}$
BM	18	43.0	3.5	(41.27, 44.73)	(40.72, 45.28)	2.110	2.785
CO	13	40.4	3.3	(38.42, 42.38)	(37.76, 43.04)	2.179	2.926
SO	17	43.1	3.2	(41.46, 44.74)	(40.94, 45.26)	2.120	2.805
SMO	14	47.1	3.2	(45.27, 48.93)	(44.66, 49.54)	2.160	2.888



The widths of the individual C.I.s are considerably wider than the standard 95% C.I. in order to achieve a 95% simultaneous C.I. for the k C.I.s.

The four C.I.'s now overlap and thus it would not appear that the means for the four regimens are different.

In STAT 642, we will learn procedures to perform the above inferences in a more efficient manner using multiple comparison and ANOVA techniques.

In many instances, the individual C.I.s are not independent. How do we construct simultaneous C.I.s?

In some settings there will be more efficient methods for deriving simultaneous C.I.s. than the method described below.

However, the Bonferroni procedure is widely used in those cases where more efficient procedures do not exist.

Bonferroni Simultaneous C.I.s for k parameters

Suppose we have $100(1-\alpha)\%$ C.I.s on k parameters $\theta_1, \ldots, \theta_k$. However, the C.I.s are not necessarily independent. A lower bound on the simultaneous coverage probability of the k C.I.s is obtained using the Bonferroni inequality:

Let A_i be the event $\{\theta_i \epsilon\, (L_i, U_i)\}$

$$\begin{aligned}
\gamma &= P\left[\theta_i \epsilon\, (L_i, U_i) \quad \text{for} \quad i = 1, \ldots, k\right] \\
&= P\left[\cap_{i=1}^{k} A_k\right] \\
&= 1 - P\left[\cup_{i=1}^{k} A_k^c\right] \\
&\geq 1 - \sum_{i=1}^{k} P\left[A_i^c\right] = 1 - \sum_{i=1}^{k} \alpha_i = 1 - k\alpha \quad \text{if} \quad \alpha_i \equiv \alpha
\end{aligned}$$

Thus, to set the level of the k individual C.I.s, let

$$\alpha = \frac{1-\gamma}{k}$$

Simultaneous inferences and C.I.s obtained using the Bonferroni inequality are often inefficient because the actual coverage probability may be much larger than the nominal coverage γ.

The following table illustrates the coverage probability needed on k individual C.I.s to obtain a $100\gamma\%$ simultaneous C.I.s for k parameters using the Bonferroni inequality:

$$1 - \alpha = 1 - \frac{1-\gamma}{k} \;\Rightarrow\; \alpha/2 = .5(1-\gamma)/k$$

k	1	2	3	4	5	$\cdots$	13	14
γ	0.9500	0.9500	0.9500	0.9500	0.9500	$\cdots$	0.9500	0.9500
$1-\alpha$	0.9500	0.9750	0.9833	0.9875	0.9900	$\cdots$	0.9962	0.9964
$\alpha/2$	0.02500	0.0125	0.008333	0.006250	0.005000	$\cdots$	0.001923	0.001786

Note, that the width of the individual C.I.s will be somewhat wider than a 0.95 C.I. in order to achieve a 95% simultaneous coverage probability.

For example, if $k = 4$, then to obtain four C.I.s having simultaneous Coverage Probability of .95, it would be necessary to construct four individual CIs have coverage probability of .9875, that is, we would use $\alpha/2 = .00625$ in place of $\alpha/2 = .025$ if we were not concerned about simultaneous coverage. Also, note that $\alpha/2$ has decreased from .006371 to .00625 due to the lack of independence between the four C.I.'s. This will result in slightly wider intervals.

Wald C.I. based on MLE's

When we obtain MLE of parameters, the R or SAS output will provide asymptotic estimates of the standard errors of the MLE's.

For example, if $\hat{\theta}$ is the MLE of θ, then we can also obtain the asymptotic standard error of $\hat{\theta}$, $\widehat{SE}(\hat{\theta})$. Using these values, an asymptotic approximate $100(1-\alpha)$ for θ is given by

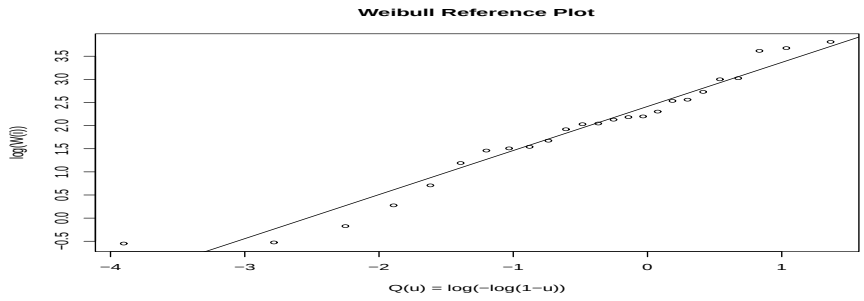
$$\hat{\theta} \pm Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}) \;=\; (\hat{\theta} - Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}),\;\; \hat{\theta} + Z_{\alpha/2} \cdot \widehat{SE}(\hat{\theta}))$$

where $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile of the $N(0,1)$ distribution.

Example: Suppose we have a random sample of n=23 ball bearings and observe the number of revolutions to failure for the 23 ball bearings:

```
17.88  28.92   33.00    41.52   42.12   45.60   48.40  51.84
51.96  54.12   55.56    67.80   68.64   68.64   68.88  84.12
93.12  98.64  105.12   105.84  127.92  128.04    173.40
```

The researcher states that from previous studies that the Weibull distribution was a good approximation to cdf of the r.v. R, Revolutions to Failure. A Weibull Probability plot and the Anderson-Darling GOF measure were applied to the data. The results are given here:



From the R output we have ADM = 0.3413 which implies from Table 5 in Handout 9 that $p-value > 0.25$.

From the p-value and the Weibull Reference Distribution plot we would conclude that the Weibull distribution provides an excellent fit to the bearing data. Next, we obtain the MLE's of the parameters, γ and α:

```
*R Code to find MLE:

library(MASS)

x <- c(
17.88 , 28.92 ,  33.00 ,   41.52 ,  42.12 ,  45.60 ,  48.40,  51.84 ,
51.96 , 54.12 ,  55.56   , 67.80 ,  68.64  , 68.64 ,  68.88 , 84.12 ,
93.12 , 98.64 , 105.12  , 105.84 , 127.92 , 128.04 , 173.40)

fitdistr(x,"weibull")

output from R code:

    shape        scale
   2.1011178   81.8324383
 ( 0.3285826) ( 8.5971353)
```

From the R code we obtain our parameter estimates:

Estimate of γ is $\hat{\gamma} = 2.1012$ with estimated standard error $\widehat{SE}(\hat{\gamma}) = .3286$

Estimate of α is $\hat{\alpha} = 81.8324$ with estimated standard error $\widehat{SE}(\hat{\alpha}) = 8.5971$

We can then construct approximate 95% C.I. for the two parameters:

$$95\% \text{ C.I. for } \gamma: \quad \hat{\gamma} \pm 1.96 \cdot \widehat{SE}(\hat{\gamma}) = 2.1012 \pm 1.96 \cdot .3286 = (1.457,\ 2.745)$$

$$95\% \text{ C.I. for } \alpha: \quad \hat{\gamma} \pm 1.96 \cdot \widehat{SE}(\hat{\alpha}) = 81.8324 \pm 1.96 \cdot 8.5971 = (73.235,\ 90.430)$$

The two intervals are fairly wide reflecting the small sample size, n=23.

We can construct similar intervals for the mean and median which are outputted from R when we use the Kaplan-Meier estimates of the survival function.

Recall the example from Handout 7:

EXAMPLE: In an experiment to determine the strength of a braided cord after weathering, the strengths of 48 pieces of cord that had been weathered for a specified length of time were investigated. The company wanted to estimate the probability that the cord would have strength of at least 53, that is, estimate S(53). Seven cords were damaged during the study which resulted in a decrease in their strength. Therefore, the study produced right censored strength values. The strengths of the remaining 41 cords were determined as shown below:

```
36.3  52.4  54.8  57.1  60.7  41.7  52.6  54.8  57.3
43.9  52.7  55.1  57.7  49.4  53.1  55.4  57.8
50.1  53.6  55.9  58.1  50.8  53.6  56.0  58.9
51.9  53.9  56.1  59.0  52.1  53.9  56.5  59.1
52.3  54.1  56.9  59.6  52.3  54.6  57.1  60.4
```

The 7 censored strength values from the damaged cords are given next:

```
26.8  29.6  33.4  35.0  40.0  41.9 42.5
```

The true strength values of the 7 cords, T_i are unobservable but we know $T_i > Y_i$, where Y_i are the observed values.

An analysis of the above data will be conducted using the following R code:

```
 library(MASS)
 library(survival)


st = c(36.3,  52.4,  54.8,  57.1,  60.7,  41.7,  52.6,  54.8, 57.3,
       43.9,  52.7,  55.1,  57.7,  49.4,  53.1,  55.4,  57.8, 50.1,
       53.6,  55.9,  58.1,  50.8,  53.6,  56.0,  58.9,  51.9, 53.9,
       56.1,  59.0,  52.1,  53.9,  56.5,  59.1,  52.3,  54.1, 56.9,
       59.6,  52.3,  54.6,  57.1,  60.4,
       26.8,  29.6,  33.4,  35.0,  40.0,  41.9,  42.5)

stcens = c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
           0,0,0,0,0,0,0)

Surv(st, stcens)

cords.surv <- survfit(Surv(st, stcens) ~ 1,conf.type="log-log")
summary(cords.surv)
print(cords.surv,print.rmean=TRUE)
```

Estimators from R Code:

```
records   n.max    n.start  events  *rmean  *se(rmean)  median  0.95LCL    0.95UCL
 48.000  48.000    48.000   41.000   54.182   0.723     54.800   53.100     56.100
* restricted mean with upper limit =  60.7
```

The above output from R displays the estimated value of the mean

$\hat{\mu} = 54.182$ with estimated standard error $\widehat{SE}(\hat{\mu}) = .723$.

From these values, we can compute an approximate 99% C.I. for the mean, $Z_{.005} = 2.576$:

$$\hat{\mu} \pm 2.576 \cdot \widehat{SE}(\hat{\mu}) \quad = \quad 54.182 \pm 2.576 \cdot .723 \quad = \quad (52.32,\ 56.04)$$

A 99% confidence interval for the median can computed in a similar fashion once we know $\widehat{SE}(\hat{median})$

The width of the 95% C.I. is $(56.1 - 53.1) = 2(1.96)\widehat{SE}(median) \ \Rightarrow \ \widehat{SE}(median) = 0.7653$

The 99% C.I. for the median is

$$54.8 \pm 2.576 \cdot .7653 \quad = \quad (52.83,\ 56.77)$$

.

.

.