Notes for Class-20: April 30

Bootstrap confidence intervals
(Empirical, Efron's, Non-parametric)

See the file notes_5_bootstrap.pdf
on Canvas.

Apologies for missing office hours!

Also I am off email etc. from 10:30 today, until Sunday pm: draft slides are uploaded.

Jie will publish the zoom recording.

Updates

- Homework 4: coming in: 32 by 9:00 pm
- Miniquiz 5: due tomorrow Friday
- Homework 5: is posted
- Lab 6:   will be posted Monday.
  - Lab6 requires the data set abalone.csv which is now available on Canvas – see under the Lab files or the Home page Labs module
- Projects: one group has reported their leader.
  - Next week – no miniquiz
  - instead updates from each student
    on their group, data set interests etc.
- Gone Friday 10:30 am to Sunday 1:00 pm
  - Jie will publish class recording.
  - Draft class slides are posted

20.1 Normal Confidence Intervals

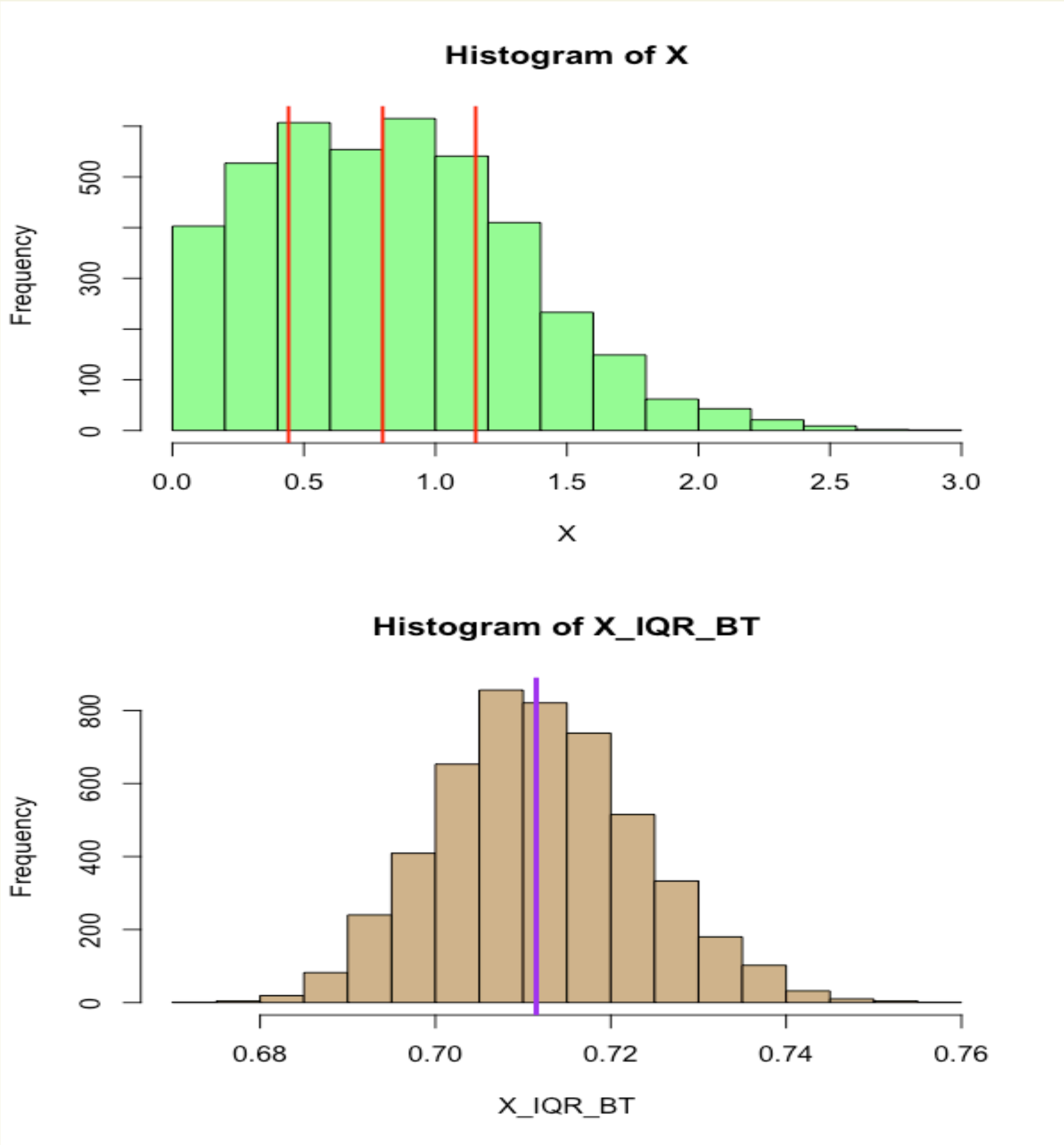
- To form confidence intervals we need assumptions
- Suppose $\widehat{\theta} = t(x_1, \ldots, x_n),$ ~N(0,τ^2) (large n?)
- Suppose bootstrap samples gives a good estimate of dsn of θ^; so SD($\theta^{*(b)}$) ≈ τ = sd of θ^. (large n?)
- Then a (1-α)-level confidence interval for θ is

$$(\widehat{\theta} - z_{1-\alpha/2}\,\widehat{\mathrm{sd}}(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2}\,\widehat{\mathrm{sd}}(\widehat{\theta}))$$

- where $z_{1-\alpha/2} = \Phi^{-1}(1-\alpha/2)$ quantile of N(0,1):
  That is $\Phi^{-1}(0.975) = 1.96$ for α=0.05.
- For example, for the median, confidence interval is

$$M_n \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathsf{Var}}_B(M_n)}.$$

Example: Abalone whole weight IQR



- X=whole weight
- n=4177
- X_IQR = 0.7155

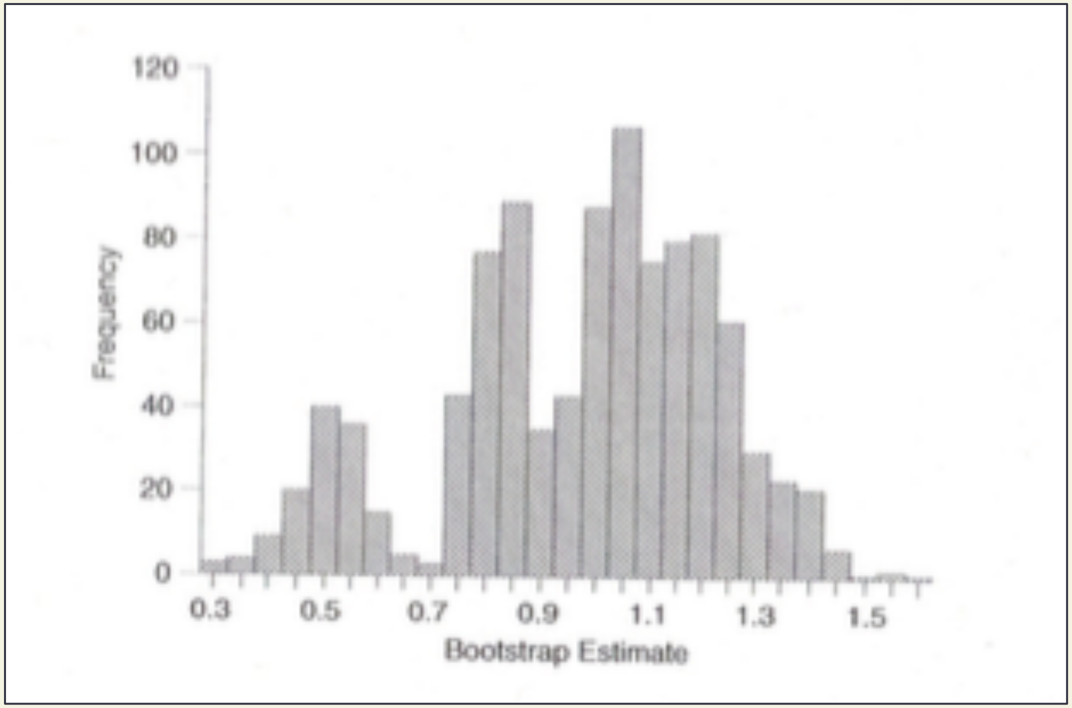
- B=5000
  X_IQR_BT values
- Sd-est = SD(X_IQR_BT) = 0.0116

- **CI for population IQR is**
- **0.7155 ± 1.96*0.0116**
  =(0.6927,0.7382)

20.2: Example: confidence limits for an SD

- Suppose we have a sample size n=20, which in fact is from an exponential, Exp(1),distribution. True SD=1
- Note, this example is for illustration only: n=20 is small, and turns out this is hard problem for bootstrap.
- Sample is (3.56, 0.69, 0.10, 1.84, 3.93, 1.25, 0.18, 1.13, 0.27, 0.50, 0.67, 0.01, 0.61, 0.82, 1.70, 0.39, 0.11, 1.20, 1.21, 0.72).
- The sample SD is: $\widehat{\theta} = \sqrt{\sum_{i=1}^{20}(x_i - \overline{x})^2/20} = 1.03.$
  Note, for illustration, we here use the MLE estimate (divide by n=20, not (n-1)=19).
- Bootstrap samples, giving SDs $\theta^{*(b)}$, b=1,..,B=1000.
- Mean $\theta^{*(b)}$ = 0.97, SD($\theta^{*(b)}$) =0.25, estimates SD(θ^).
- Confidence interval 1.03 ± 1.96×0.25= (0.54, 1.52)

Histogram of the 1000 bootstrap SDs



Note the original sample has 2 large values— the lower mode are basically from the bootstrap samples that contain neither.

- n=20 is small sample, and bootstrap SDs is not well approximated by a Normal distribution.
- The histogram of the 1000 bootstrap SDs is not encouraging

20.3 Bias corrected Normal CI

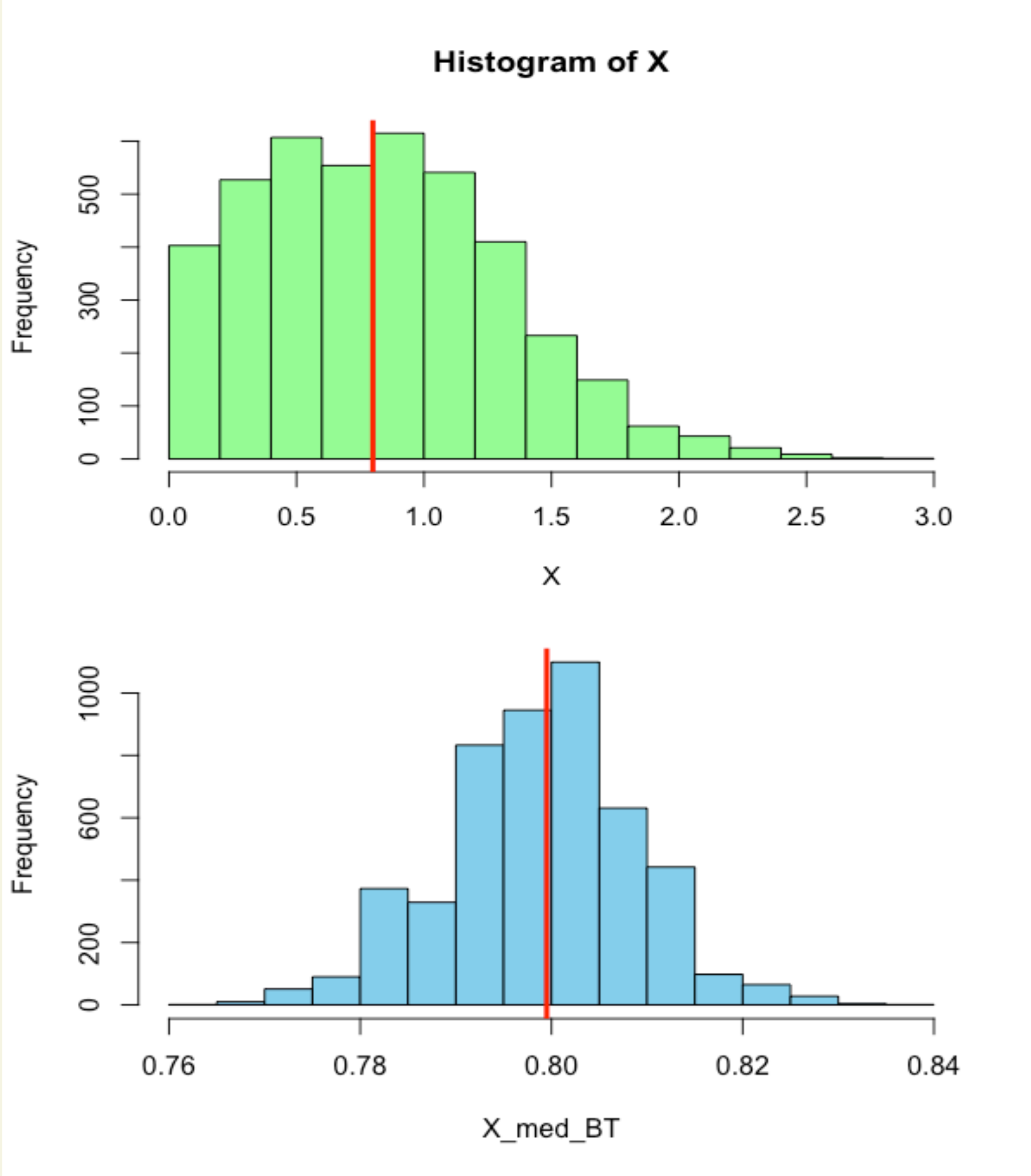
- Mean $\theta^{*(b)}$=0.97, so estimated bias=0.97−1.03 =−0.06.
- That is, the lower values in the bootstrap resamples suggest the sample estimate 1.03 may be too low.
- The true value could be estimated as 1.03+0.06=1.09
- 95% Confidence Interval for population SD becomes

  $(1.09 \pm 1.96 \times 0.25) = (0.60, 1.58)$
- These intervals (bias corrected or not), look OK for the true SD=1, but this was perhaps good luck.
- Repeat the process 1000 times: i.e sample n=20 from Exp(1) and for each make a bootstrap CI with B=1000.
- Mean of bias-corrected estimates = 0.98. [True=1✓]
- However, only 73% of the supposedly 95% bias-corrected CI covered true value 1.

20.4 Quantile-based confidence interval

- Suppose now $g(\widehat{\theta}) \sim \mathcal{N}(g(\theta), 1)$, where g is unknown, but increasing function of θ.
- Then $I_C = (g(\widehat{\theta}) - z_{1-\alpha/2}, g(\widehat{\theta}) + z_{1-\alpha/2})$ is a C=100(1-α)% confidence interval for g(θ).
- Then Bootstrap resample estimates $g(\theta^{*(b)})$ should be approx $N(g(\widehat{\theta}), 1)$, so fraction C should fall in I_C
- So a C-level interval for g(θ) is the central C propn of the bootstrap estimates $g(\theta^{*(b)})$.
- g monotonic, so ordering of $g(\theta^{*(b)})$ is same as for $\theta^{*(b)}$
- So bootstrap confidence interval for θ is $(\theta^*_{\alpha/2}, \theta^*_{(1-\alpha/2)})$ where these are the empirical quantiles of the bootstrap samples e.g.0.025,0.975 for 95%

Abalone whole weight example



- X=whole weight
- n=4177
- Median = X_med = 0.7995

- B=5000
- Histogram is counts of 5000 bootstrap medians X_med_BT
- Line is at sample median 0.7995.
- **CI for true population median would be central 95% of this**

20.5 Alternate (Hall's) quantile based CI

- Instead of using the bootstrap values $\theta^{*(b)}$ to approximate the distribution of θ^, we use the distribution of differences $\varepsilon_b = (\theta^{*(b)} - \hat{\theta})$ to approximate the distribution of $(\hat{\theta} - \theta)$.
- That is we choose ε_L and ε_H such that

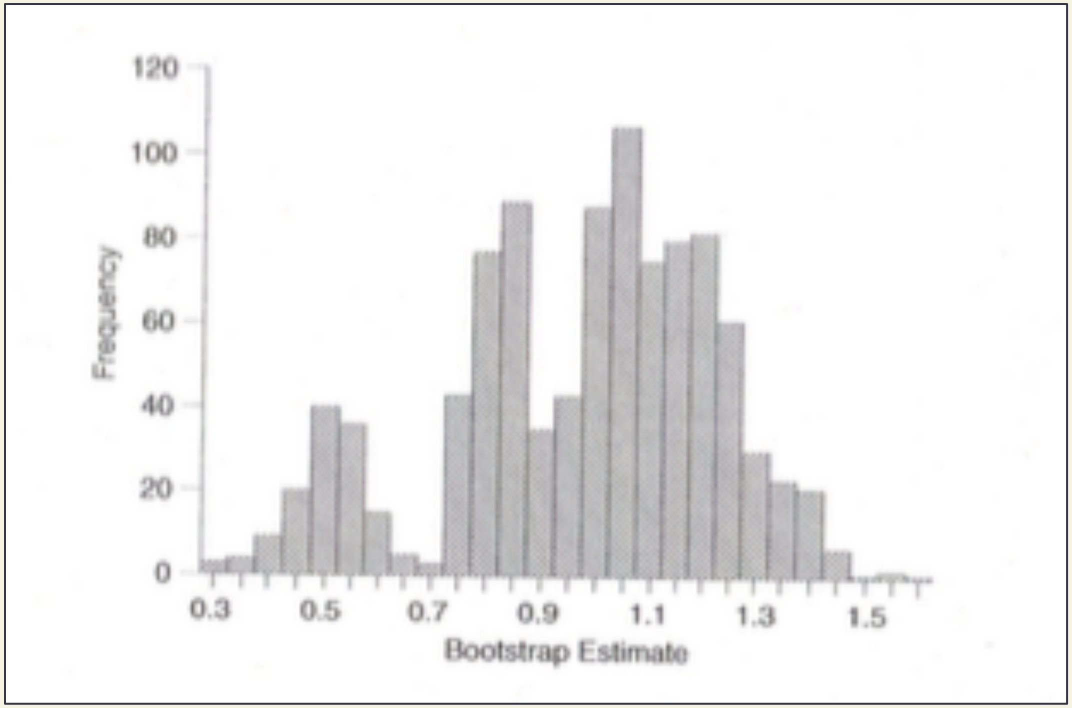
$$P(\varepsilon_L < (\theta^{*(b)} - \widehat{\theta}) < \varepsilon_H)) = 1 - \alpha$$

- Choose $\varepsilon_L = \theta^*_{\alpha/2} - \hat{\theta}$, and $\varepsilon_H = \theta^*_{1-\alpha/2} - \hat{\theta}$
- Then we can say a (1-α)-level

  CI for θ is given by $P(\varepsilon_L < (\widehat{\theta} - \theta) < \varepsilon_H)) = 1 - \alpha$
- This leads to (1-α)-level confidence interval for θ:

  $= (\hat{\theta} - \varepsilon_H, \hat{\theta} - \varepsilon_L) =$ $(2\widehat{\theta} - \theta^*_{1-\alpha/2}, 2\widehat{\theta} - \theta^*_{\alpha/2})$

Histogram of the 1000 bootstrap SDs



Note the original sample has 2 large values— the lower mode are basically from the bootstrap samples that contain neither.

- n=20 is small sample, and bootstrap SDs is not well approximated by a Normal distribution.
- The histogram of the 1000 bootstrap SDs is not encouraging

Example: the 1000 Bootstrap SDs

- Using the basic 0.025 and 0.975 quantiles of the 1000 bootstrap SDs, we have an interval 0.44 to 1.40.
- Using the alternate (Hall's) method the limits are
  $2\times1.03 - 1.40 = 0.66$ and $2\times1.03 - 0.44 = 1.62$
- Neither is good. Repeating the experiment 1000 times;
  - Generate sample size n=20 of Exp(1): compute sample SD
  - Do 1000 Bootstrap samples, compute 1000 $SD^{(b)}$s
  - Construct nominal 95% quantile-based interval
- For basic type, true value (=1) included only 65.9% of intervals
- For alternate type, true value (=1) included in 72.7% of intervals

20.6 Bias-corrected quantile intervals

- Suppose now $g(\widehat{\theta}) \sim \mathcal{N}(g(\theta) - z_0, 1)$, where g is unknown increasing function, and z_0 unknown bias correction.
- Note we cannot just redefine $g_1(\theta) = g(\theta)+z_0$, as then mean becomes $g(\theta)$, not $g_1(\theta)$.
- Now: $\mathrm{P}(g(\widehat{\theta}) + z_0 - z_{1-\alpha/2} < g(\theta) < g(\widehat{\theta}) + z_0 + z_{1-\alpha/2}) = (1-\alpha)$

  and we want to choose z_0.
- Since g is monotone $\mathrm{P}(\widehat{\theta} > \theta) = \mathrm{P}(g(\widehat{\theta}) > g(\theta)) = \mathrm{P}(Z > z_0)$ where $Z \sim N(0,1)$ leading to an estimate of z_0:

$$\widehat{z}_0 = \Phi^{-1}\left[1 - \frac{1}{B}\sum_{b=1}^{B} \mathbf{1}_{\widehat{\theta}_b > \widehat{\theta}}\right]$$

- Consider the upper CI limit:

$$\mathrm{P}(g(\theta^*_B) < g(\widehat{\theta})+z_0+z_{1-\alpha/2}) = \mathrm{P}(g(\theta^*_B)-g(\widehat{\theta})+z_0 < z_0+z_{1-\alpha/2}+z_0)$$

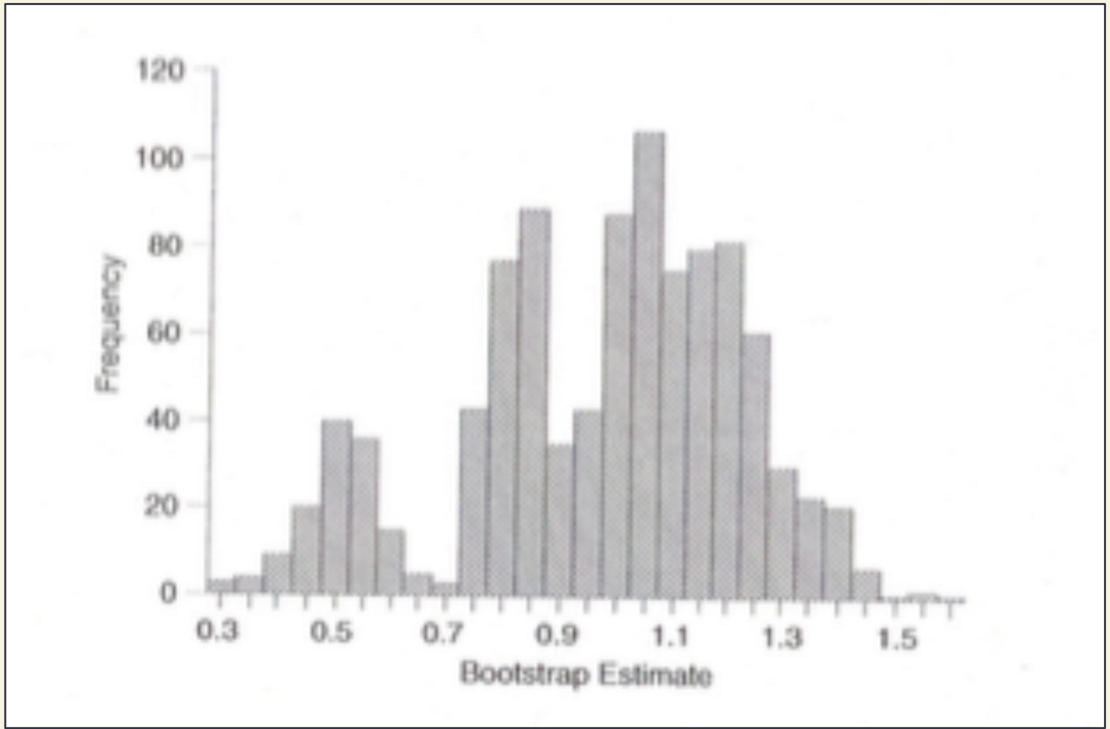
$$= \mathrm{P}(Z < 2z_0+z_{1-\alpha/2}) = \Phi(2z_0+z_{1-\alpha/2}).$$

Ctd: bias-corrected quantile intervals

- Thus the upper limit of the CI is the empirical α_2 quantile of the bootstrap distribution where

$$\alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2})$$

- Similarly the lower limit is the empirical α_1 quantile, where $\alpha_1 = \Phi(2\hat{z}_0 - z_{1-\alpha/2})$
- Remember, here are the 1000 bootstrap SDs from sample n=20 from Exp(1) (True SD=1)



20.7: Example of the SDs from n=20

- The estimate of the SD from the sample was 1.03
- Of the 1000 bootstrap SDs, only 400 were > 1.03:
  estimate P(θ^>θ) by P($θ^*_B$ > θ^) = 400/1000 = 0.4
- Hence $z_0 = \Phi^{-1}(1\text{-}0.4) = 0.25$.
- $$\alpha_1 = \Phi(2\hat{z}_0 - z_{1-\alpha/2})$$ = Φ(0.5 – 1.96) = 0.072
  $$\alpha_2 = \Phi(2\hat{z}_0 + z_{1-\alpha/2})$$ = Φ(0.5 + 1.96) = 0.993
- The bias corrected interval cuts off 7% at the lower end and <1% at the top end.
- Note if the median bootstrap value is the sample value, z_0=0, and we get back to the symmetric interval.

Blank slide

- Improves display of last slide of class