Bootstrap: Introduction to Why and How

C. Durso

CSV Workshop, April 2015

Distributions of Statistics

Estimates of distributions of some statistics are tractable for some population distributions. For these, inference is straightforward.

- Mean of an iid Normal sample: $\frac{\bar{x}-\mu}{\sqrt{\frac{\Sigma(x_i-\bar{x})^2}{n-1}}}$ has a Student t-distribution with n-1 degrees of freedom.
- Variance of an iid Normal sample: $\frac{\Sigma(x_i-\bar{x})^2}{n-1}$ has a χ^2 distribution with n-1 degrees of freedom.
- There are others...

The list doesn't go on and on.

What if Your Statistic Isn't on the List?

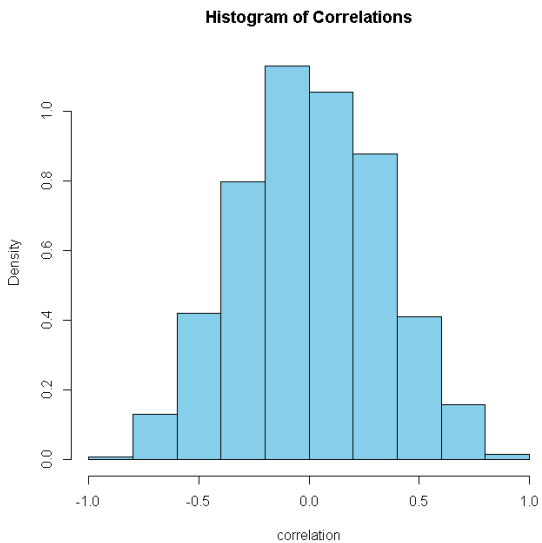
- Population distribution is unknown.
- Relation between the population distribution and the distribution of the statistic is unknown.

Simulate!

Correlation Example

simulation not strictly necessary

If $x_1, x_2 \ldots x_n$ and $y_1, y_2 \ldots y_n$ are iid samples from a $Normal(0,1)$, the standard normal distribution, what is the distribution of the sample correlation r? Look and see, for n=10.



95% interval: (-0.62, 0.65)
proportion of correlations between -0.3 and 0.3: 0.62

Exact Correlation Distribution

check

Turns out, $t = r\sqrt{\frac{n-2}{1-r^2}}$ is Student t with n-2 degrees of freedom. (So $r = \frac{t}{\sqrt{n-2+t^2}}$)

Approximate symmetric 95% interval for r with n=10: $t \in (-2.3, 2.3)$, so $r \in (-0.63, 0.63)$

Probability that r lies between -0.3 and 0.3:

$Pr\left(t \in \left(-0.3\sqrt{\frac{10-2}{1-0.3^2}}, 0.3\sqrt{\frac{10-2}{1-0.3^2}}\right)\right) \approx 0.60$

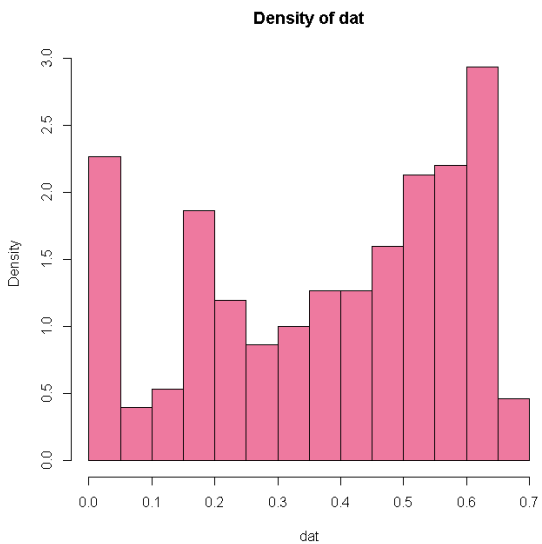
The exact version is much more complicated for bivariate normal data.

Unknown Population Distribution

Data may have an unknown distribution

No, really?

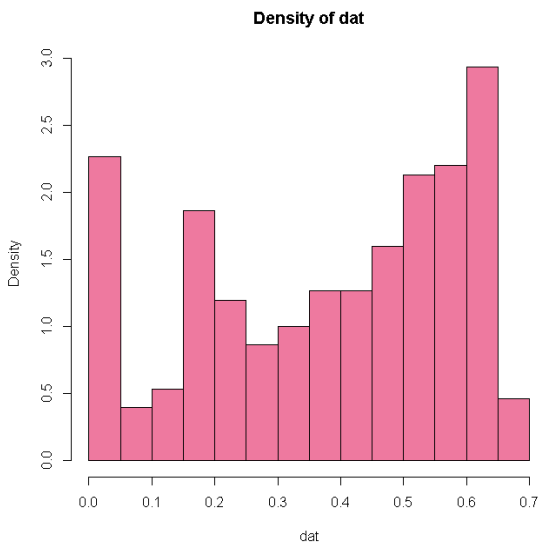
1. Fit a distribution from a flexible parametrized family. (another time)
2. Use the sample population. (this time)



Unknown Population Distribution

Data may have an unknown distribution
No, really?

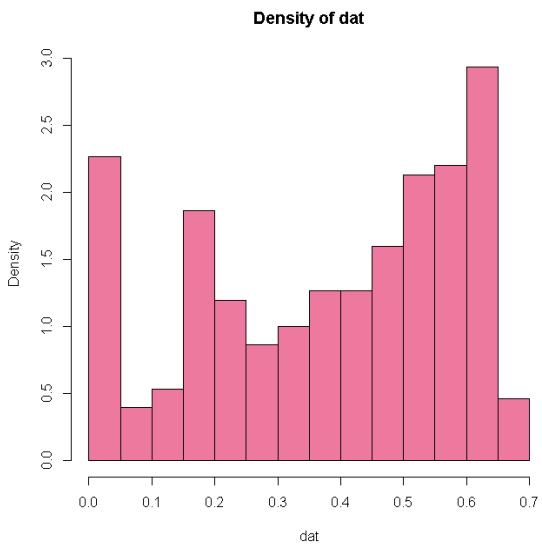
1. Fit a distribution from a flexible parametrized family. (another time)
2. Use the sample population. (this time)



Unknown Population Distribution

Data may have an unknown distribution
No, really?

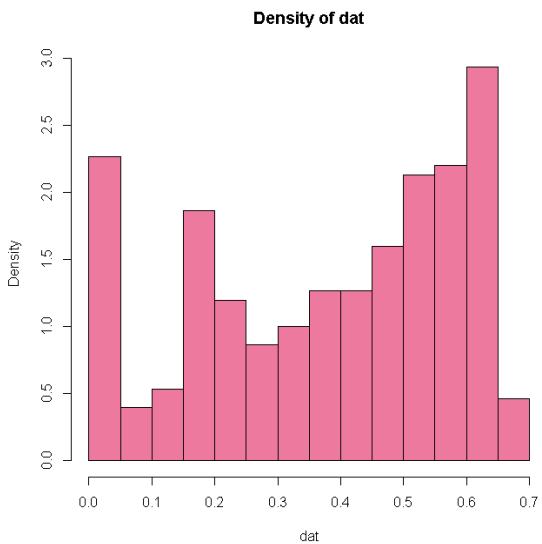
1. Fit a distribution from a flexible parametrized family. (another time)
2. Use the sample population. (this time)



Unknown Population Distribution

Data may have an unknown distribution
No, really?

1. Fit a distribution from a flexible parametrized family. (another time)
2. Use the sample population. (this time)



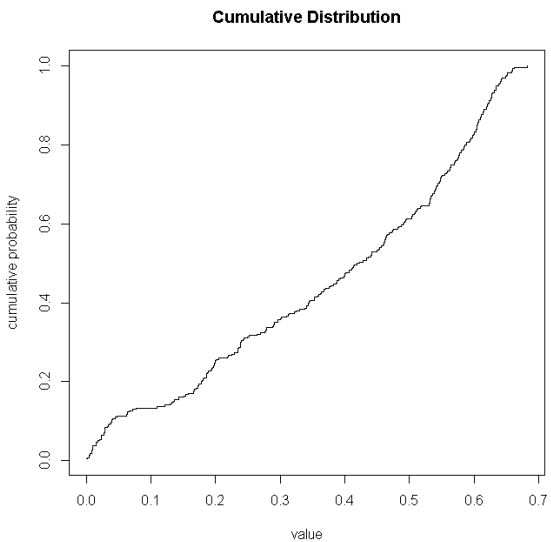
Empirical Distribution

based on sample $\{x_1, x_2, \ldots x_n\}$

Random variable X: $Pr(X = x) =$(count of x in sample) / (size of sample).

Expected value: $E[X] = \frac{1}{n}\sum x_i = \bar{x}$

Variance: $Var[X] = \frac{1}{n}\sum (x_i - \bar{x})^2$



Estimate Population Interquartile Range

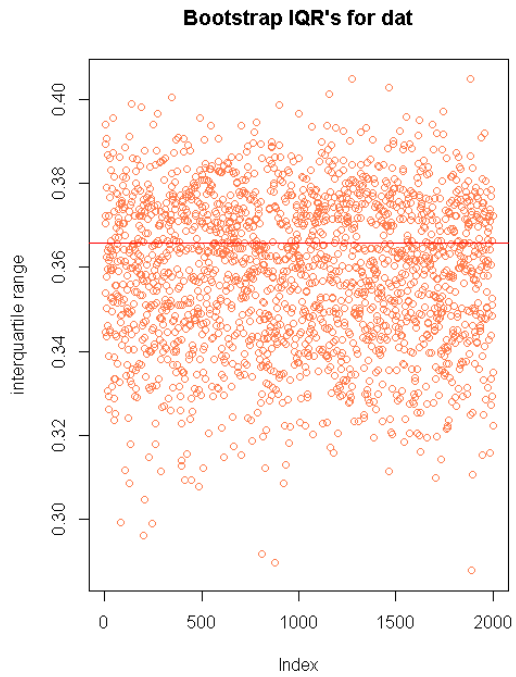
sample of size 300 from dat's population

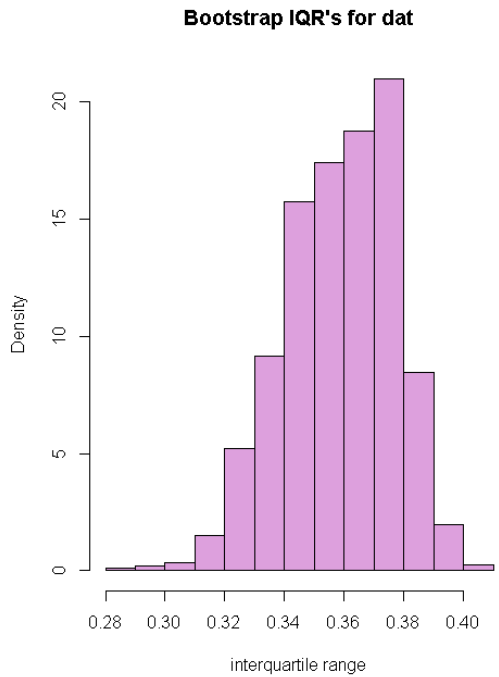
Interquartile range for dat:0.37 plus or minus what?

Simulate!

Wait. We don't have a population distribution. Approximate it by the empirical distribution.

1. Sample the data with replacement 300 times.
2. Calculate the interquartile range of the result.
3. Repeat 2000 times, or so.





Terms

Definition

bootstrap sample: Given a data set $\mathbf{x} = \{x_1, x_2, ...x_n\}$, a bootstrap sample from $\mathbf{x}$ is a set $\{x_{i_1}, x_{i_2}, ...x_{i_n}\}$ where $\{i_1, i_2, ...i_n\}$ is a sample with replacement from $\{1, 2, ...n\}$.

Definition

bootstrap values of the parameter : collection of values of the parameter evaluated on a set of bootstrap samples

95% Interval

Estimate a 95% confidence interval for the interquartile range:

- Use symmetric 95% interval of simulation interquartile ranges
- Get $(0.32, 0.39)$
- first order accurate: $Prob\left\{\theta < \hat{\theta}_{low}\right\} \doteq .025 + \frac{c_{low}}{\sqrt{n}}$,

  $Prob\left\{\theta > \hat{\theta}_{high}\right\} \doteq .025 + \frac{c_{high}}{\sqrt{n}}$

  - Efron, B., Tibshirani, R.B.,(1998) *An Introduction to the Bootstrap*. CRC , p. 187

Better 95% Interval

BC_a, bias corrected and accelerated

Adjust percentiles for bounds by

- $\hat{z}_0$ measure of difference between original estimate and median of the bootstrap values
- $\hat{a}$ measure of the rate of change of the standard error of the estimated value of the parameter $\hat{\theta}$ with repect to the population value of the parameter θ

Get values α_1 and α_2 to use in place of .025 and .975, or, generally α and $1-\alpha$.

Gain: second order accurate: $Prob\left\{\theta < \hat{\theta}_{low}\right\} \doteq .025 + \frac{c_{low}}{n}$,

$Prob\left\{\theta > \hat{\theta}_{high}\right\} \doteq .025 + \frac{c_{high}}{n}$ (Efron, B., Tibshirani, R.B.,(1998))

Symbol Definitions

- Denote the standard normal cumulative distribution function by Φ.
- Let $z^{(\beta)}$ be the value with $\Phi\left(z^{(\beta)}\right) = \beta$.
- Let B be the number of bootstrap samples.
- Let $\hat{\theta}_{(i)}$ be the estimate of θ based on all the observations except the i^{th}.
- Let $\hat{\theta}_{(.)}$ be the mean of the $\hat{\theta}_{(i)}$'s.
- $\hat{\theta*}$is the collection of bootstrap estimates of θ
- Intended coverage is $1-2\alpha$

Festival of Formulae

- $\hat{z}_0 = \Phi^{-1}\left(\frac{\text{proportion of bootstrap values less than sample estimate } \hat{\theta}}{B}\right)$
- $\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^3}{6\left\{\sum_{i=1}^{n}\left(\hat{\theta}_{(\cdot)} - \hat{\theta}_{(i)}\right)^2\right\}^{\frac{3}{2}}}$
- $\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(\alpha)}\right)}\right)$
- $\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{a}\left(\hat{z}_0 + z^{(1-\alpha)}\right)}\right)$
- $BC_a \text{interval for } 1 - 2\alpha = \left(\hat{\theta*}^{(\alpha_1)}, \hat{\theta*}^{(\alpha_2)}\right)$

For this Data

One estimate is $(0.3339, 0.3978)$. With another random seed, $(0.3312, 0.3970)$.
Compare to percentile version $(0.3213, 0.3893)$.