Statistical Consulting Topics

The Bootstrap...

“The bootstrap is a computer-based method for assigning measures of accuracy to statistical estimates.” (Efron and Tibshirani, 1998.)

- What do we do when our distributional assumptions about our model are not met?

- The assumptions are needed to give us...
  - valid standard errors
  - valid confidence intervals
  - valid hypothesis tests and $p$-values
• Consider inference on a mean $\mu$ with the estimator $\bar{X}$ from an $i.i.d$ random sample.
  
  – $\bar{X}$ is a random variable

  – Define the statistic: $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

  – If the population from which we are drawing is normally distributed, we have: $T \sim t_{n-1}$ which we use to...
    * form valid $100(1 - \alpha)\%$ CI’s on $\mu$
    * perform $\alpha$-level hypothesis tests on $\mu$

  – Without normality, we can not assume $T$ has this distribution (its dist’n is unknown)

  – How can we do inference?
• The **Bootstrap** method can be used for estimating a sampling distribution when the theoretical sampling distribution is unknown.

Recall, a **sampling distribution** is...

> the probability distribution of a statistic.

Examples (true under certain conditions):

1. \( \bar{X} \sim N(\mu, \frac{\sigma^2}{n}) \)

2. \( \hat{\beta} \sim N(\beta, V(\hat{\beta})) \)

These sampling distributions allow us to perform hypothesis tests and form confidence intervals on parameters of interest.

If ‘conditions’ are not met, we can still perform statistical inference by using the **bootstrap**.
Bootstrapping:  
Example for inference on $\rho$  
(Population correlation)

- Average values for GPA and LSAT scores for students admitted to $n=15$ Law Schools in 1973 (a random sampling of law schools).

<table>
<thead>
<tr>
<th>School</th>
<th>LSAT</th>
<th>GPA</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>576</td>
<td>3.39</td>
</tr>
<tr>
<td>2</td>
<td>635</td>
<td>3.30</td>
</tr>
<tr>
<td>3</td>
<td>558</td>
<td>2.81</td>
</tr>
<tr>
<td>4</td>
<td>579</td>
<td>3.03</td>
</tr>
<tr>
<td>5</td>
<td>666</td>
<td>3.44</td>
</tr>
<tr>
<td>6</td>
<td>580</td>
<td>3.07</td>
</tr>
<tr>
<td>7</td>
<td>555</td>
<td>3.00</td>
</tr>
<tr>
<td>8</td>
<td>661</td>
<td>3.43</td>
</tr>
<tr>
<td>9</td>
<td>651</td>
<td>3.36</td>
</tr>
<tr>
<td>10</td>
<td>605</td>
<td>3.13</td>
</tr>
<tr>
<td>11</td>
<td>653</td>
<td>3.12</td>
</tr>
<tr>
<td>12</td>
<td>575</td>
<td>2.74</td>
</tr>
<tr>
<td>13</td>
<td>545</td>
<td>2.76</td>
</tr>
<tr>
<td>14</td>
<td>572</td>
<td>2.88</td>
</tr>
<tr>
<td>15</td>
<td>594</td>
<td>2.96</td>
</tr>
</tbody>
</table>

- Can we make a confidence interval for $\rho$?
- Point estimate for $\rho$ is the sample correlation $r$:

$$ r = 0.7766 $$

- Classical inference (like confidence intervals) on $\rho$ depends on $X$ and $Y$ having a bivariate normal distribution.

- In the above sample, there are a few outliers suggesting this assumption may be violated.

- We’ll use the bootstrap approach to do statistical inference instead.
– Sampling $n=15$ new cases as $(x_i, y_i)$ with replacement from my original data, I create a new bootstrapped data set and calculate $r$ which I will label as $r_1^\ast$.

– Repeating this process 1000 times provides an empirical ‘sampling distribution’ for the estimator $r$ (empirical means based on the observed data).

\[ r_1^\ast, r_2^\ast, r_3^\ast, \ldots, r_{1000}^\ast \]

The distribution of the above values gives us an idea of the variability of our estimator $r$.

– We will assume these $n = 15$ observations is a representative sample from the population of all law schools (the assumption we make).
The distribution of $r_1^*, r_2^*, \ldots, r_{1000}^*$ is the empirical sampling distribution of our estimator $r$.

We can use it to make a 95% empirical confidence interval for $\rho$.

```r
> quantile(bootstrapped.r.values, 0.025, type=3)
2.5%
0.4419053

> quantile(bootstrapped.r.values, 0.975, type=3)
97.5%
0.9623332
```
– We were able to create a CI without any assumptions on distribution, i.e. nonparametrically (very useful in many situations).

– This only works if the original sample is representative of the original population.

– Recall what sampling variability of an estimator is... BEFORE we collect our data, the estimator is a random variable because it’s value depends on the sample chosen.

– The bootstrap method uses resampling to get a handle on this variability (since we can’t get at it theoretically because our assumptions weren’t met).

– We should resample from the $n$ observations in the same manner as how the original data was sampled (here, we had a simple random sample).
Classical test of $H_0 : \rho = 0$ (option 1)

- Built on the assumption that $X$ and $Y$ follow a bivariate normal distribution with $\theta = (\mu_X, \mu_Y, \sigma^2_X, \sigma^2_Y, \rho)$.

- $\hat{\rho} = r$, and

$$r = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}}$$

- Test of $H_0 : \rho = 0$

$$T = r\sqrt{n - 2}/\sqrt{1 - r^2}$$

Under $H_0$, $T \sim t_{n-2}$

This is the same $t$-test as $H_0 : \beta_1 = 0$ in simple linear regression.
Classical confidence interval (CI) for $\rho$

- $r$ is obviously not normally distributed.
- Fisher's $z$-transformation of $r$:

$$Z_r = \frac{1}{2} \log\left(\frac{1+r}{1-r}\right)$$

- $Z_r$ is approx $N\left(\frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n-3}\right)$

- Transformation creates approximate normality, and variance independent of $\rho$
  
  - $100(1-\alpha)$% CI for $Z_r$ is
    $$Z_r \pm z_{\alpha/2} \sqrt{\frac{1}{n-3}}$$

  - To get CI for $\rho$, invert transformation of lower limit and upper limit for $Z_r$ using
    $$r = \frac{(e^{2Z_r} - 1)}{(e^{2Z_r} + 1)}$$
Classical test of $H_0 : \rho = \rho_o$ (option 2)

- Based on Fisher’s $Z_r$

- Test statistic:

$$Z = \sqrt{n - 3}(Z_r - \frac{1}{2} \log\left(\frac{1 + \rho_o}{1 - \rho_o}\right))$$

Under $H_o$, $Z \sim N(0, 1)$
Compare CIs from Fisher's $Z_r$ and Bootstrap

• 95% CI based on Fisher’s $Z_r$:

If our observed data is bivariate normal, then

$$Z_r \text{ is approx } N\left(\frac{1}{n-3}, \frac{1}{n-3}\right)$$

and our 95% CI for $\frac{1}{2} \log\left(\frac{1+\rho}{1-\rho}\right)$ is

$$(0.4710, 1.6026)$$

and our 95% CI for $\rho$ is

$$(0.4399, 0.9221).$$

• 95% CI based on Bootstrap (shown earlier):

$$(0.4419, 0.9623).$$
If our observed data is bivariate normal, then the distribution of transformed bootstrapped correlations $Z^*_{r,i}$s should be approximately normally distributed where

$$Z^*_{r,i} = \frac{1}{2} \log \left( \frac{1 + r^*_i}{1 - r^*_i} \right).$$

Histogram of $Z.r.bootstrapped$

$\hat{\mu}_{Z^*_r} = 1.1112$ and $\hat{\sigma}^2_{Z^*_r} = 0.1435$

And $\sigma^2_{Z^*_r,Fisher} = \frac{1}{n-3} = 0.0833$
The variance of the bootstrap distribution is slightly larger than expected by Fisher.

The $Z_{r,i}^*$ distribution is visually non-normal (though only slightly).

This suggests that the bootstrap may be a better option for inference than using the Fisher transformation which assumed the bi-variate normal.

- Is this a way to check assumptions in complex models? (Classical CI vs. Bootstrap CI)
  - I’ve seen it used in that way (pharmaceutical drug modeling, for instance).
  - Not really checking the full ‘distribution’ of estimators.
  - Perhaps as a way to say it definitely DOESN’T match assumptions.
• Some comments...

– The type of bootstrapping described here (case resampling) considers the regressors \(X\)'s to be random, not fixed.

– There are numerous types of bootstrapping methods that relate to how you create the bootstrapped data sets, such as, parametric bootstrap, semi-parametric bootstrap, wild bootstrap, or moving block bootstrap.

– The intention is for a bootstrapped data set to represent another new hypothetical data set that could have just as easily been observed as the original data set.

– If you have known correlation in your data (non-independence), this must be accounted for in your bootstrapping procedure.
– There are bias correction methods for estimates and confidence intervals.

– The Bootstrap can fail:

* Initial random sample not representative of population

* Sample too small, empirical $CDF$ too ‘jagged’ (might be fixed with smoothing)

* Application of the Bootstrap did not adequately replicate the random process that produced the original sample
Changepoint example:

In streams, the sediment (sand, gravel, rocks, etc.) that moves along the bottom (bed) of the stream is called the bedload. The bedload is distinct from suspended load (in the water) and wash load (near the top of the water).

How quickly the bedload moves in a stream (bedload transport, kg/s) is dependent on the flow of the water (discharge, m$^3$/s) and a variety of other factors.
The bedload transport has been described as occurring in phases, with a changepoint occurring at the flow where the stream transitions from phase I to phase II.

“The fitted line for less-than-breakpoint flows had a lower slope with less variance due to the fact that bedload at these discharges consist primarily of small quantities of sand-sized materials. In contrast, the fitted line for flows greater than the breakpoint had a significantly steeper slope and more variability in transport rates due to the physical breakup of the armor layer, the availability of subsurface material, and subsequent changes in both the size and volume of sediment in transport[3].”
Simulated data:

- \( n = 15 \) streams were randomly selected from the population of streams.
- At each stream, 40 observation were taken at random discharge levels.

- We’re interested in estimating the population mean changepoint \( \gamma \).
Applying the Bootstrap method for inference:

1. Calculate a $\hat{\gamma}$ from the observed data as
   \[
   \hat{\gamma}_{obs} = \frac{\sum_{i=1}^{15} \hat{\gamma}_i}{15}.
   \]

2. Generate $B$ bootstrap samples and calculate $\hat{\gamma}_b$ for each bootstrap sample $b = 1, \ldots, B$.

Two thoughts...

(a) Bootstrapping the 15 estimated change-points ($B=500$).
(b) Bootstrapping the 15 creeks and bootstrapping the observations within each creek.

95% Bootstrap Confidence intervals:
(a) (4.214, 4.811)
(b) (4.162, 4.744)

95% Confidence interval based on normality:
(4.183, 4.844)
• References:

