Bootstrapping
18.05 Spring 2018
Agenda

- Leftover from 5/2: binomial confidence intervals
- Bootstrap terminology
- Bootstrap principle
- Empirical bootstrap
- Parametric bootstrap
Board question: exact binomial confidence interval

Use this table of binomial(8,θ) probabilities to:

1. **Color** the (two-sided) rejection region with significance level 0.10 for each value of θ.
2. Given x = 7, find the 90% confidence interval for θ.
3. Repeat for x = 4.

<table>
<thead>
<tr>
<th>θ \ x</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>0.430</td>
<td>0.383</td>
<td>0.149</td>
<td>0.033</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>.3</td>
<td>0.058</td>
<td>0.198</td>
<td>0.296</td>
<td>0.254</td>
<td>0.136</td>
<td>0.047</td>
<td>0.010</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>.5</td>
<td>0.004</td>
<td>0.031</td>
<td>0.109</td>
<td>0.219</td>
<td>0.273</td>
<td>0.219</td>
<td>0.109</td>
<td>0.031</td>
<td>0.004</td>
</tr>
<tr>
<td>.7</td>
<td>0.000</td>
<td>0.001</td>
<td>0.010</td>
<td>0.047</td>
<td>0.136</td>
<td>0.254</td>
<td>0.296</td>
<td>0.198</td>
<td>0.058</td>
</tr>
<tr>
<td>.9</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.005</td>
<td>0.033</td>
<td>0.149</td>
<td>0.383</td>
<td>0.430</td>
</tr>
</tbody>
</table>
Solution

For each $\theta$, the non-rejection region is blue, the rejection region is red. In each row, the rejection region has probability at most $\alpha = 0.10$.

<table>
<thead>
<tr>
<th>$\theta \backslash x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.430</td>
<td>0.383</td>
<td>0.149</td>
<td>0.033</td>
<td>0.005</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>0.3</td>
<td>0.058</td>
<td>0.198</td>
<td>0.296</td>
<td>0.254</td>
<td>0.136</td>
<td>0.047</td>
<td>0.010</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td>0.5</td>
<td>0.004</td>
<td>0.031</td>
<td>0.109</td>
<td>0.219</td>
<td>0.273</td>
<td>0.219</td>
<td>0.109</td>
<td>0.031</td>
<td>0.004</td>
</tr>
<tr>
<td>0.7</td>
<td>0.000</td>
<td>0.001</td>
<td>0.010</td>
<td>0.047</td>
<td>0.136</td>
<td>0.254</td>
<td>0.296</td>
<td>0.198</td>
<td>0.058</td>
</tr>
<tr>
<td>0.9</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.005</td>
<td>0.033</td>
<td>0.149</td>
<td>0.383</td>
<td>0.430</td>
</tr>
</tbody>
</table>

For $x = 7$ the 90% confidence interval for $\theta$ is $[0.7, 0.9]$. These are the values of $\theta$ we wouldn’t reject as null hypotheses. They are the blue entries in the $x = 7$ column.

For $x = 4$ the 90% confidence interval for $\theta$ is $[0.3, 0.7]$. 
Board question: polling 20 instead of 8

Use this table of \( \text{pbinom}(x, 20, \theta) \) to:

1. **Color** the (two-sided) rejection region with significance level 0.05 for each value of \( \theta \).

2. **Given** \( x = 3 \), find the 95% confidence interval for \( \theta \).

3. Repeat for \( x = 10 \).

<table>
<thead>
<tr>
<th>( \theta )</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.122</td>
<td>.392</td>
<td>.677</td>
<td>.867</td>
<td>.957</td>
<td>.989</td>
<td>.998</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>.2</td>
<td>.012</td>
<td>.069</td>
<td>.206</td>
<td>.411</td>
<td>.630</td>
<td>.804</td>
<td>.913</td>
<td>.968</td>
<td>.990</td>
<td>.997</td>
<td>.999</td>
</tr>
<tr>
<td>.3</td>
<td>.001</td>
<td>.008</td>
<td>.036</td>
<td>.107</td>
<td>.238</td>
<td>.416</td>
<td>.608</td>
<td>.772</td>
<td>.887</td>
<td>.952</td>
<td>.983</td>
</tr>
<tr>
<td>.4</td>
<td>0</td>
<td>.001</td>
<td>.004</td>
<td>.016</td>
<td>.051</td>
<td>.126</td>
<td>.25</td>
<td>.416</td>
<td>.596</td>
<td>.755</td>
<td>.872</td>
</tr>
<tr>
<td>.5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.001</td>
<td>.006</td>
<td>.021</td>
<td>.058</td>
<td>.132</td>
<td>.252</td>
<td>.412</td>
</tr>
<tr>
<td>.6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.002</td>
<td>.006</td>
<td>.021</td>
<td>.056</td>
<td>.128</td>
<td>.245</td>
</tr>
<tr>
<td>.7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.001</td>
<td>.005</td>
<td>.017</td>
<td>.048</td>
</tr>
<tr>
<td>.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>.001</td>
<td>.003</td>
<td></td>
</tr>
<tr>
<td>.9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Solution

For each $\theta$, the non-rejection region is blue, the rejection region is red. In each row, the rejection region has probability at most $\alpha = 0.05$.

<table>
<thead>
<tr>
<th>$\theta \backslash x$</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>.1</td>
<td>.122</td>
<td>.392</td>
<td>.677</td>
<td>.867</td>
<td>.957</td>
<td>.989</td>
<td>.998</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>.2</td>
<td>.012</td>
<td>.069</td>
<td>.206</td>
<td>.411</td>
<td>.630</td>
<td>.804</td>
<td>.913</td>
<td>.968</td>
<td>.990</td>
<td>.997</td>
<td>.999</td>
</tr>
<tr>
<td>.3</td>
<td>.001</td>
<td>.008</td>
<td>.036</td>
<td>.107</td>
<td>.238</td>
<td>.416</td>
<td>.608</td>
<td>.772</td>
<td>.887</td>
<td>.952</td>
<td>.983</td>
</tr>
<tr>
<td>.4</td>
<td>.000</td>
<td>.001</td>
<td>.004</td>
<td>.016</td>
<td>.051</td>
<td>.126</td>
<td>.250</td>
<td>.416</td>
<td>.596</td>
<td>.755</td>
<td>.872</td>
</tr>
<tr>
<td>.5</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.006</td>
<td>.021</td>
<td>.058</td>
<td>.132</td>
<td>.252</td>
<td>.412</td>
<td>.588</td>
</tr>
<tr>
<td>.6</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.002</td>
<td>.006</td>
<td>.021</td>
<td>.056</td>
<td>.128</td>
<td>.245</td>
</tr>
<tr>
<td>.7</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.005</td>
<td>.017</td>
<td>.048</td>
</tr>
<tr>
<td>.8</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.001</td>
<td>.003</td>
</tr>
<tr>
<td>.9</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
<td>.000</td>
</tr>
</tbody>
</table>

For $x = 3$ the 95% confidence interval for $\theta$ is $[0.1, 0.3]$. These are the values of $\theta$ we wouldn’t reject as null hypotheses.

For $x = 10$ the 95% confidence interval for $\theta$ is $[0.3, 0.7]$.

*Conservative normal* confidence interval for $\theta$ is

$$x/20 \pm 1/\sqrt{20} = x/20 \pm 0.22$$

Exact confidence intervals computed here are a bit smaller.
Empirical distribution of data

Data: $x_1, x_2, \ldots, x_n$ (independent)

**Example 1.** Data: 1, 2, 2, 3, 8, 8, 8.

<table>
<thead>
<tr>
<th>$x^*$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>$p^<em>(x^</em>)$</td>
<td>1/7</td>
<td>2/7</td>
<td>1/7</td>
<td>3/7</td>
</tr>
</tbody>
</table>

**Example 2.**

The true and empirical distribution are approximately equal.
Resampling

- Sample (size 6): 1 2 1 5 1 12

- **Resample (size $m$)**: Randomly choose $m$ samples with replacement from the original sample.

- Resample probabilities = empirical distribution: $P(1) = 1/2$, $P(2) = 1/6$ etc.

- E.g. resample (size 10): 5 1 1 1 12 1 2 1 1 5

- A **bootstrap (re)sample** is always the same size as the original sample:

- Bootstrap sample (size 6): 5 1 1 1 12 1
Bootstrap principle for the mean

- Data $x_1, x_2, \ldots, x_n \sim F$ with true mean $\mu$.
- $F^* =$ empirical distribution (resampling distribution).
- $x_1^*, x_2^*, \ldots, x_n^*$ resample same size data

Bootstrap Principle: (really holds for any statistic)

1. $F^* \approx F$ computed from resample; $\bar{x}^*$ for mean.
2. $\delta^* = \bar{x}^* - \bar{x} \approx \bar{x} - \mu =$ variation of $\bar{x}$.
3. Critical values:
   $$\delta_{1-\alpha/2}^* \leq \bar{x}^* - \bar{x} \leq \delta_{\alpha/2}^*$$
   except for $\alpha$ extreme cases.
4. Bootstrap confidence interval for $\mu$ is
   $$\bar{x} - \delta_{\alpha/2}^* \leq \mu \leq \bar{x} - \delta_{1-\alpha/2}^*$$
Empirical bootstrap confidence intervals

Use the data to estimate the variation of estimates based on the data!

- **Data:** $x_1, \ldots, x_n$ drawn from a distribution $F$.
- **Estimate a feature** $\theta$ of $F$ by a statistic $\hat{\theta}$.
- **Generate many bootstrap samples** $x_1^*, \ldots, x_n^*$.
- **Compute the statistic** $\theta^*$ for each bootstrap sample.
- **Compute the bootstrap difference** $\delta^* = \theta^* - \hat{\theta}$.
- **Use quantiles of** $\delta^*$ **to approximate quantiles of** $\delta = \hat{\theta} - \theta$.
- **Construct a confidence interval** $[\hat{\theta} - \delta^*_{\alpha/2}, \hat{\theta} - \delta^*_{1-\alpha/2}]$
  (By $\delta^*_{\alpha/2}$ we mean the $\alpha/2$ critical value.)
Consider finding bootstrap confidence intervals for

I. the mean  
II. the median  
III. 47th percentile.

Which is easiest to find?

A. I  
B. II  
C. III  
D. I and II  
E. II and III  
F. I and III  
G. I and II and III

answer: G. The program is essentially the same for all three statistics. All that needs to change is the code for computing the specific statistic.
Board question

Data: 3 8 1 8 3 3

Bootstrap samples (each column is one bootstrap trial):

\[
\begin{align*}
&8 8 1 8 3 8 3 1 \\
&1 3 3 1 3 8 3 3 \\
&3 1 1 8 1 3 3 8 \\
&8 1 3 1 3 3 8 8 \\
&3 3 1 8 8 3 8 3 \\
&3 8 8 3 8 3 1 1
\end{align*}
\]

Compute a bootstrap 80% confidence interval for the mean.

Compute a bootstrap 80% confidence interval for the median.
Solution: mean

\[ \bar{x} = 4.33 \]

\[ \bar{x}^*: \quad 4.33, 4.00, 2.83, 4.83, 4.33, 4.67, 4.33, 4.00 \]

\[ \delta^*: \quad 0.00, -0.33, -1.50, 0.50, 0.00, 0.33, 0.00, -0.33 \]

Sorted

\[ \delta^*: \quad -1.50, -0.33, -0.33, 0.00, 0.00, 0.00, 0.33, 0.50 \]

So, \( \delta_{0.9}^* = -1.50, \quad \delta_{0.1}^* = 0.37 \).

(For \( \delta_{0.1}^* \) we interpolated between the top two values –there are other reasonable choices. In R see the `quantile()` function.)

80% bootstrap CI for mean: \[ [\bar{x} - 0.37, \bar{x} + 1.50] = [3.97, 5.83] \]
Solution: median

\[ x_{0.5} = \text{median}(x) = 3 \]

\[ x_{0.5}^*: \quad 3.0, 3.0, 2.0, 5.5, 3.0, 3.0, 3.0, 3.0 \]

\[ \delta^*: \quad 0.0, 0.0, -1.0, 2.5, 0.0, 0.0, 0.0, 0.0 \]

Sorted

\[ \delta^*: \quad -1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.5 \]

So, \( \delta_{0.9}^* = -1.0, \quad \delta_{0.1}^* = 0.5 \).

(For \( \delta_{0.1}^* \) we interpolated between the top two values – there are other reasonable choices. In R see the \texttt{quantile()} function.)

80% bootstrap CI for median: \[ [\bar{x} - 0.5, \bar{x} + 1.0] = [2.5, 4.0] \]
Empirical bootstrapping in R

```r
x = c(30,37,36,43,42,43,43,46,41,42)  # original sample
n = length(x)  # sample size
xbar = mean(x)  # sample mean
nboot = 5000  # number of bootstrap samples to use

# Generate nboot empirical samples of size n
# and organize in a matrix
tmpdata = sample(x,n*nboot, replace=TRUE)
bootstrapsample = matrix(tmpdata, nrow=n, ncol=nboot)

# Compute bootstrap means xbar* and differences delta*
xbarstar = colMeans(bootstrapsample)
deltastar = xbarstar - xbar

# Find the .1 and .9 quantiles and make
# the bootstrap 80% confidence interval
.ci = quantile(deltastar, c(.1,.9))
ci = xbar - c(d[2], d[1])```
Parametric bootstrapping

Use the estimated parameter to estimate the variation of estimates of the parameter!

- Data: \( x_1, \ldots, x_n \) drawn from a parametric distribution \( F(\theta) \).
- Estimate \( \theta \) by a statistic \( \hat{\theta} \).
- **Generate many bootstrap samples from \( F(\hat{\theta}) \).**
- Compute the statistic \( \theta^* \) for each bootstrap sample.
- Compute the bootstrap difference \( \delta^* = \theta^* - \hat{\theta} \).
- Use crit values of \( \delta^* \) to approximate crit values of \( \delta = \hat{\theta} - \theta \).
- Set a bootstrap confidence interval \( [\hat{\theta} - \delta^*_{1-\alpha/2}, \hat{\theta} - \delta^*_{\alpha/2}] \)