# 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,$\theta$) probabilities to:

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

| $\theta \backslash x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| .1 | 0.430 | 0.383 | 0.149 | 0.033 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| .3 | 0.058 | 0.198 | 0.296 | 0.254 | 0.136 | 0.047 | 0.010 | 0.001 | 0.000 |
| .5 | 0.004 | 0.031 | 0.109 | 0.219 | 0.273 | 0.219 | 0.109 | 0.031 | 0.004 |
| .7 | 0.000 | 0.001 | 0.010 | 0.047 | 0.136 | 0.254 | 0.296 | 0.198 | 0.058 |
| .9 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.033 | 0.149 | 0.383 | 0.430 |

# 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 $\alpha = 0.10$.

| $\theta \backslash x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|---|---|
| .1 | 0.430 | 0.383 | 0.149 | 0.033 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 |
| .3 | 0.058 | 0.198 | 0.296 | 0.254 | 0.136 | 0.047 | 0.010 | 0.001 | 0.000 |
| .5 | 0.004 | 0.031 | 0.109 | 0.219 | 0.273 | 0.219 | 0.109 | 0.031 | 0.004 |
| .7 | 0.000 | 0.001 | 0.010 | 0.047 | 0.136 | 0.254 | 0.296 | 0.198 | 0.058 |
| .9 | 0.000 | 0.000 | 0.000 | 0.000 | 0.005 | 0.033 | 0.149 | 0.383 | 0.430 |

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 `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 $x = 10$.

| $\theta \backslash x$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| .1 | .122 | .392 | .677 | .867 | .957 | .989 | .998 | 1 | 1 | 1 | 1 |
| .2 | .012 | .069 | .206 | .411 | .630 | .804 | .913 | .968 | .990 | .997 | .999 |
| .3 | .001 | .008 | .036 | .107 | .238 | .416 | .608 | .772 | .887 | .952 | .983 |
| .4 | 0 | .001 | .004 | .016 | .051 | .126 | .25 | .416 | .596 | .755 | .872 |
| .5 | 0 | 0 | 0 | .001 | .006 | .021 | .058 | .132 | .252 | .412 | .588 |
| .6 | 0 | 0 | 0 | 0 | 0 | .002 | .006 | .021 | .056 | .128 | .245 |
| .7 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .001 | .005 | .017 | .048 |
| .8 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | .001 | .003 |
| .9 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# 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 $\alpha = 0.05$.

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

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.

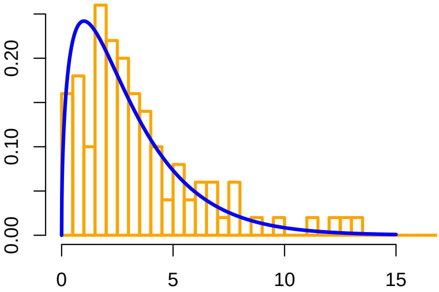
# Empirical distribution of data

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

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

| $x^*$ | 1 | 2 | 3 | 8 |
|---|---|---|---|---|
| $p^*(x^*)$ | $1/7$ | $2/7$ | $1/7$ | $3/7$ |

**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; $\overline{x}^*$ for mean.
2. $\delta^* = \overline{x}^* - \overline{x} \approx \overline{x} - \mu =$ variation of $\overline{x}$.
3. Critical values:
$$\delta^*_{1-\alpha/2} \le \overline{x}^* - \overline{x} \le \delta^*_{\alpha/2}$$
except for $\alpha$ extreme cases.
4. Bootstrap confidence interval for $\mu$ is
$$\boxed{\overline{x} - \delta^*_{\alpha/2} \le \mu \le \overline{x} - \delta^*_{1-\alpha/2}}$$

# 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**.)

# Concept question

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

<u>**answer:**</u> **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):

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 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 |

Compute a bootstrap 80% confidence interval for the mean.

Compute a bootstrap 80% confidence interval for the median.

# Solution: mean

$\bar{x} = 4.33$

$\bar{x}^*$: `4.33, 4.00, 2.83, 4.83, 4.33, 4.67, 4.33, 4.00`

$\delta^*$: `0.00, -0.33, -1.50, 0.50, 0.00, 0.33, 0.00, -0.33`

Sorted
$\delta^*$: `-1.50, -0.33, -0.33, 0.00, 0.00, 0.00, 0.33, 0.50`

So, $\delta^*_{0.9} = -1.50$, $\delta^*_{0.1} = 0.37$.

(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

$x_{0.5} = \text{median}(x) \;=\; 3$

$x^*_{0.5}$: `3.0, 3.0, 2.0, 5.5, 3.0, 3.0, 3.0, 3.0`

$\delta^*$: `0.0, 0.0, -1.0, 2.5, 0.0, 0.0, 0.0, 0.0`

Sorted
$\delta^*$: `-1.0, 0.0, 0.0, 0.0, 0.0, 0.0, 0.0, 2.5`

So, $\delta^*_{0.9} = -1.0$, $\delta^*_{0.1} = 0.5$.

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

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

# Empirical bootstrapping in 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 $[\hat{\theta} - \delta^*_{\alpha/2}, \hat{\theta} - \delta^*_{1-\alpha/2}]$