# Example 1

95% CI for $\mu_X - \mu_Y$:

The data below show the sugar content (as a % of weight) of several national brands of children's and adults' cereals.

$X$ = sugar content in a typical children's cereal — $\mu_X = E[X]$

**Children's cereals**: 40.3, 55, 45.7, 43.3, 50.3, 45.9, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.5

$Y$ = sugar content in a typical adults' cereal — $\mu_Y = E[Y]$

**Adults' cereals**: 20, 30.2, 2.2, 7.5, 4.4, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4

(a) Is it reasonable to assume that each sample comes from a normal distribution?

using R — look at boxplot — NO, may be OK for children's cereals but not for adults' cereals

(b) Can the variances of the two distributions be assumed to be equal? Justify your answer.

(c) Compute an appropriate 95% confidence interval for difference in mean sugar contents of the two cereal types. What assumptions did you make, if any, to construct the CI?

95% CI for $\mu_x - \mu_y$ = $[32.5, 40.8]$

↑ plausible values

Clearly: (a) since the entire CI is above, we can conclude that $\mu_x > \mu_y$

(b) plausible values for $\mu_x - \mu_y$: 32.5 to 40.8

Assumption: normality which may not justified, esp. for the second sample. But given that the diff. in the two dist. is so clear-cut, the conclusions are still valid.

(d) What do you conclude on the basis of your answer in (c)? Can we say that children's cereals have more sugar on average than adult cereals? If yes, by how much? Justify your answers.

Yes, see the previous prob.

All calculation done using R code given in feb22.R file on elearning.

# Example 3

Consider the dataset stored in the file bp.txt. This dataset contains one measurement of systolic blood pressure (in mmHg) made by each of two methods — a finger method and an arm method — from the same 200 patients.

(a) Perform an exploratory analysis of the data by examining the distributions of the measurements from the two methods using boxplots. Comment on what you see. Do the two distributions seem similar? Justify your answer.

saw in class

see R code for calculation & results

(b) Use histograms and QQ plots to examine the shapes of the two distributions. Comment on what you see. Does the assumption of normality seem reasonable? Justify your answer.

discussed in class

(c) Construct an appropriate 95% confidence interval for the difference in the means of the two methods. Interpret your results. Can we conclude that the two methods have identical means? Justify your answer. What assumptions, if any, did you make to construct the interval? Do the assumptions seem to hold?

95% CI for $\mu_x - \mu_y$: $[-6.3, -2.3]$

arm - finger

normality for diff. seemed reasonable but is not really needed since $n = 200$ is large

$\Rightarrow$ Since the entire interval is below zero, we have $\mu_x - \mu_y < 0 \Rightarrow \mu_x < \mu_y$

and also plausible values for $\mu_x - \mu_y$ is b/w $-6.3$ to $-2.3$ mmHg.

# Bootstrap

$F(x) = P[X \le x]$

**Set up:** Data $X_1, \ldots, X_n$ — i.i.d. as $X$ with population cdf $F$ (which is not completely known). — focus in one-sample problems.

**Parameter of interest:** $\theta$, estimated by $\hat{\theta}$

**Examples:** Mean, variance, median, quantiles, etc.

**Issue:** Need to get the *sampling distribution* of $\hat{\theta}$ so that we can compute, e.g., standard error of $\hat{\theta}$, or confidence interval for $\theta$?

**Q:** Why not use the methods that we have learnt?

| $\hat{\theta}$ | sampling distribution |
|---|---|
| • $\bar{X}$ | - $N\left[\mu, \frac{\sigma^2}{n}\right]$ if pop. is Normal |
| | - $\approx N\left[\mu, \frac{\sigma^2}{n}\right]$ if $n$ is large |
| • $\hat{\theta}_{MLE}$ | - $\approx N\left[\theta, \hat{I}^{-1}\right]$ if $n$ is large. |

- Sample median
- sample variance
- sample quantile

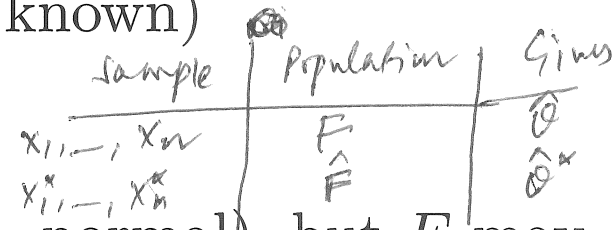
} — complicated.

# Basics

**Bootstrap:** A simulation based technique that allows us to approximate the sampling distribution of $\hat{\theta}$. Assumes large $n$, but its value needed for validity of bootstrap is typically less than that for the usual large-sample procedure.

**Original sample:** $X_1, \ldots, X_n \sim$ i.i.d. with cdf $F$

**Bootstrap (re)sample:** $X_1^*, \ldots, X_n^* \sim$ i.i.d. with cdf $\hat{F}$, where $\hat{F}$ = estimated cdf (which is completely known)




**Parametric bootstrap:**

- Functional form of $F$ is known (e.g., normal), but $F$ may depend on unknown parameter $\theta$.
- $\hat{F}$ is same as $F$ but with $\theta$ replaced by its MLE $\hat{\theta}$. In other words, $\hat{F}$ is the cdf of the fitted model.
- Ex: $F = N(\mu, \sigma^2)$, $\hat{F} = N[\bar{X}, \hat{\sigma}^2 = \frac{(n-1)s^2}{n}]$
- Often easy to simulate i.i.d. draws $X_1^*, \ldots, X_n^*$ from $\hat{F}$.

$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

fitted model

## Nonparametric bootstrap:

$I(A) = \begin{cases} 1, & \text{if } A \text{ occurs} \\ 0, & \text{otherwise} \end{cases}$

- Functional form of $F$ is unknown.
- $\hat{F}$ = empirical cdf, where

Suppose: $x_1, x_2, \ldots, x_n$ — obs. data

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) = \text{prop. of obs. that are} \le x.$$

$\hat{F}(x)$ is a step fn. that changes at $x_1 \to x_n$ and step size $= \frac{1}{n}$

- Think of $\hat{F}$ as a discrete distribution that assigns $1/n$ probability to each of the sample observations, $X_1, \ldots, X_n$.
- Get $X_1^*, \ldots, X_n^*$ by sampling $n$ times **with replacement** from $X_1, \ldots, X_n$.

consider a rv $X^*$ that takes $x_1, \ldots, x_n$ as possible values, each with equal prob. $\frac{1}{n}$ — CDF of $X^*$ is $\hat{F}$.

Need to get $x_1^*, x_2^*, \ldots, x_n^*$ as a random sample from $\hat{F}$ — use sample fn. in R.

**Bootstrap distribution of estimator $\hat{\theta}$ of $\theta$:**

**Original sample:** $X_1, \ldots, X_n$ — gives $\hat{\theta}$

- Simulate a large number $b$ of *bootstrap resamples*, and compute $\hat{\theta}^* = g(\hat{F}^*)$ from each resample. exactly the way we compute $\hat{\theta}$ from the original sample

| sample | Est. |
|---|---|
| $x_1, x_2, \ldots, x_n$ | $\hat{\theta}$ |

| Resample # | $\hat{\theta}^*$ |
|---|---|
| $x_{11}^*, x_{12}^*, \ldots, x_{1n}^*$ | $\hat{\theta}_1^*$ |
| $x_{21}^*, x_{22}^*, \ldots, x_{2n}^*$ | $\hat{\theta}_2^*$ |
| $\vdots$ | $\vdots$ |
| $x_{b1}^*, x_{b2}^*, \ldots, x_{bn}^*$ | $\hat{\theta}_b^*$ |

b draws.

- This process gives a large number of draws, $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$.
- These draws are coming from the *bootstrap distribution* of $\hat{\theta}$.
- How to see this distribution? Histogram, boxplot etc.
- It approximates the *sampling distribution* of $\hat{\theta}$.
- Use the draws $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$ to estimate features of sampling distribution of $\hat{\theta}$ that may be of interest. — essentially, using a Monte Carlo method to estimate features of a dist.

**Estimating a feature $\eta$ of distribution of $\hat{\theta}$:**

- Get a large number $b$ of draws, $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$.
- $\hat{\eta}^*$ = same feature computed from these draws.

} THIS is Monte Carlo method

**Ex 1:** $\eta = E(\hat{\theta})$. $\hat{\eta}^* =$ $\frac{1}{b}\sum_{k=1}^{b} \hat{\theta}_k^* = \bar{\hat{\theta}}^*$

**Ex 2:** $\eta = var(\hat{\theta})$. $\hat{\eta}^* =$ $\frac{1}{b-1}\sum_{k=1}^{b} (\hat{\theta}_k^* - \bar{\hat{\theta}}^*)^2$ $\Rightarrow SE(\hat{\theta}) = SD$ of the draws

**Ex 3:** $\eta =$ bias of $\hat{\theta} = E(\hat{\theta}) - \theta$. $\hat{\eta}^* = \bar{\hat{\theta}}^* - \hat{\theta}$

**Ex 4:** $\eta = \alpha$-th quantile of $\hat{\theta}$. $\hat{\eta}^* =$ $\alpha$-th sample quantile of the draws $\hat{\theta}_1^*, \ldots, \hat{\theta}_b^*$

$= \hat{\theta}_{(b\alpha)} \rightarrow (b\alpha)$-th ordered Obs.

will use: $\hat{\theta}_{((b+1)\alpha)}$

**Ex 5:** $\eta = \alpha$-th quantile of $(\hat{\theta} - \theta)$. $\hat{\eta}^* =$

$= (\alpha\text{-th quantile of } \hat{\theta}) - \theta$

$\Rightarrow \hat{\eta}^* = \hat{\theta}_{(b\alpha)} - \hat{\theta}$

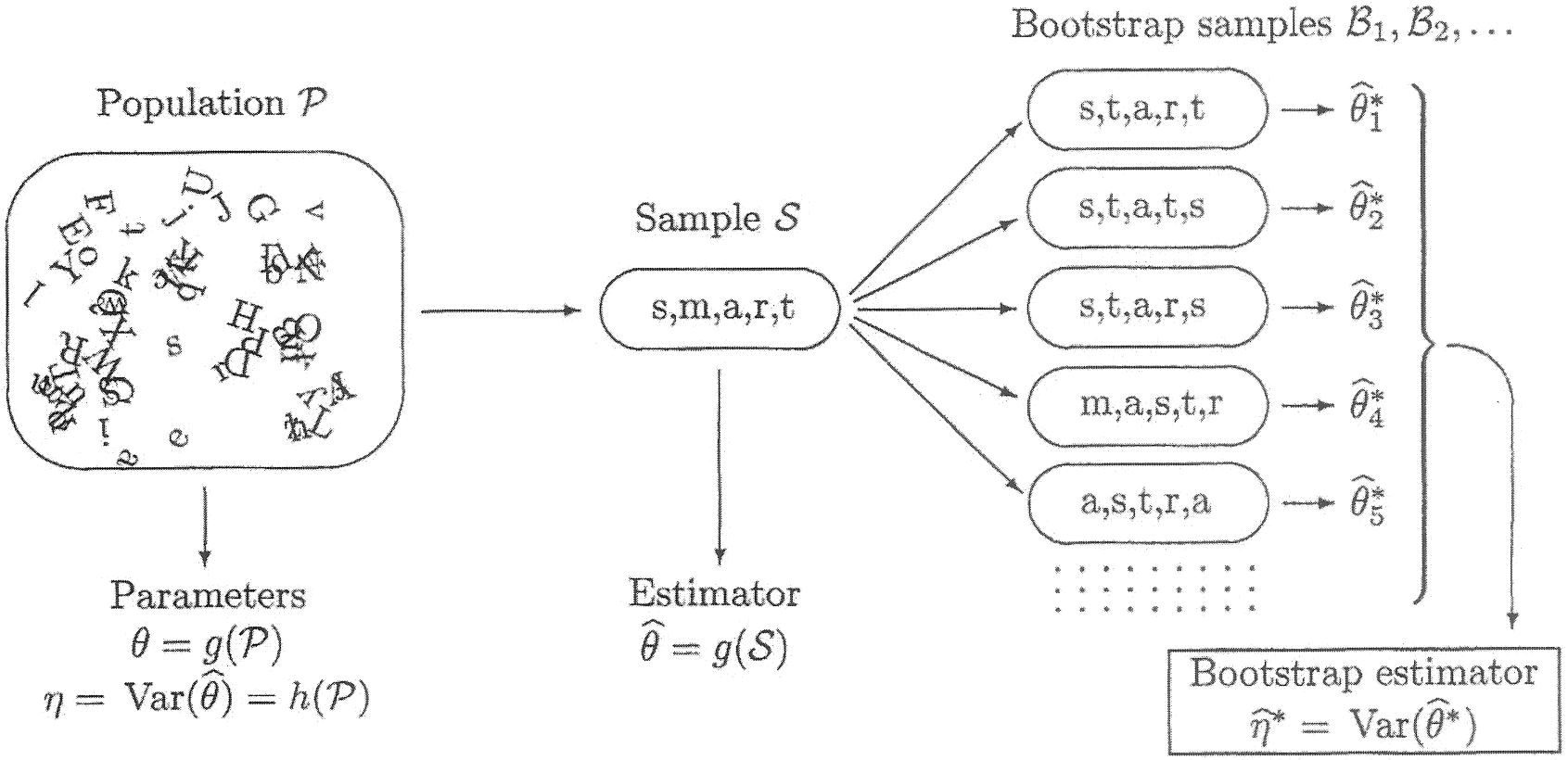



FIGURE 10.6: *Bootstrap procedure estimates* $\eta = \mathrm{Var}(\widehat{\theta})$ *by the variance of* $\widehat{\theta}^*_i$*'s, obtained from bootstrap samples.*