Statistical modeling of usual intake via short-term dietary assessment instruments

Kevin W. Dodd

Biometry Research Group, Division of Cancer Prevention

NIH NATIONAL CANCER INSTITUTE

June 19, 2018

Outline

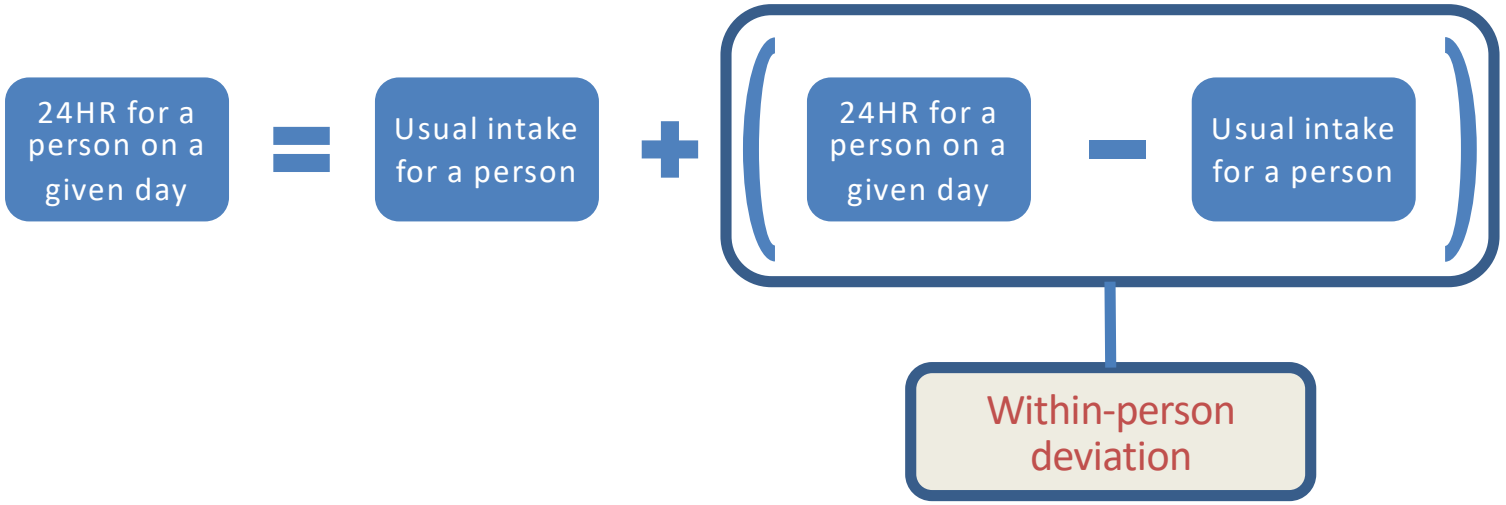
- Modeling
  - Basic (one-part)
  - Two-part (for episodically consumed dietary components)
- Subpopulations and the role of covariates
- Survey design

What does “modeling” entail?

- A way of filling in gaps in information using statistical techniques
- In this case, pooling limited information from sampled individuals
- Requires assumptions

Foundation of the model

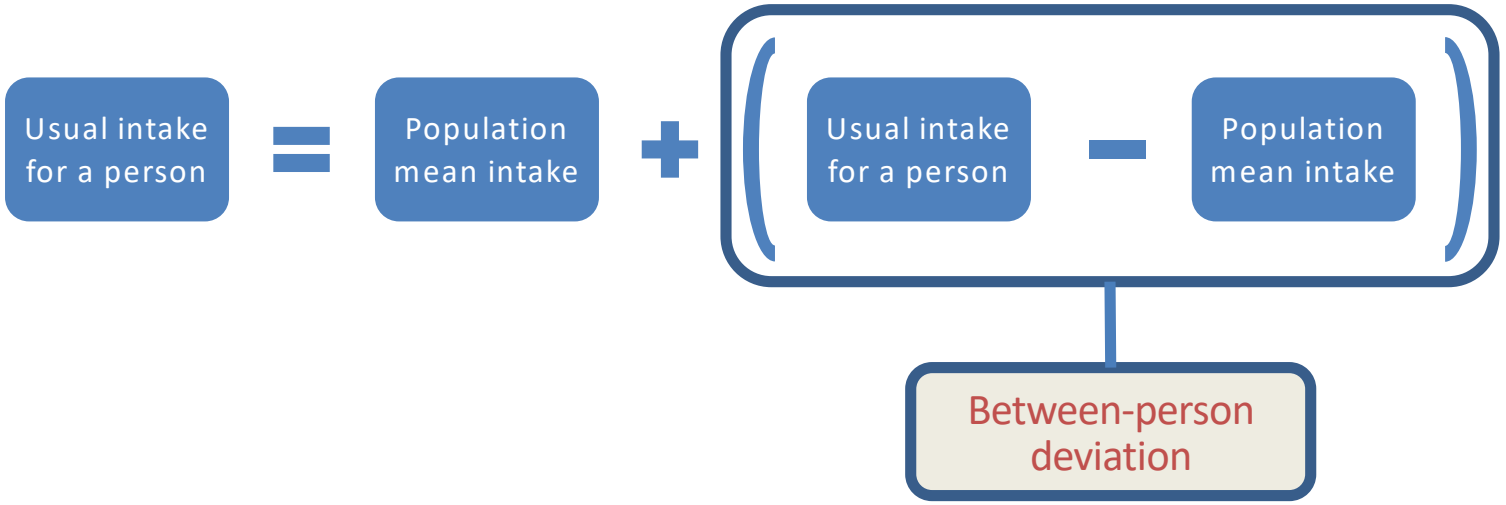
- Each recall is a usual intake plus a deviation



- Within-person deviations cancel out across days

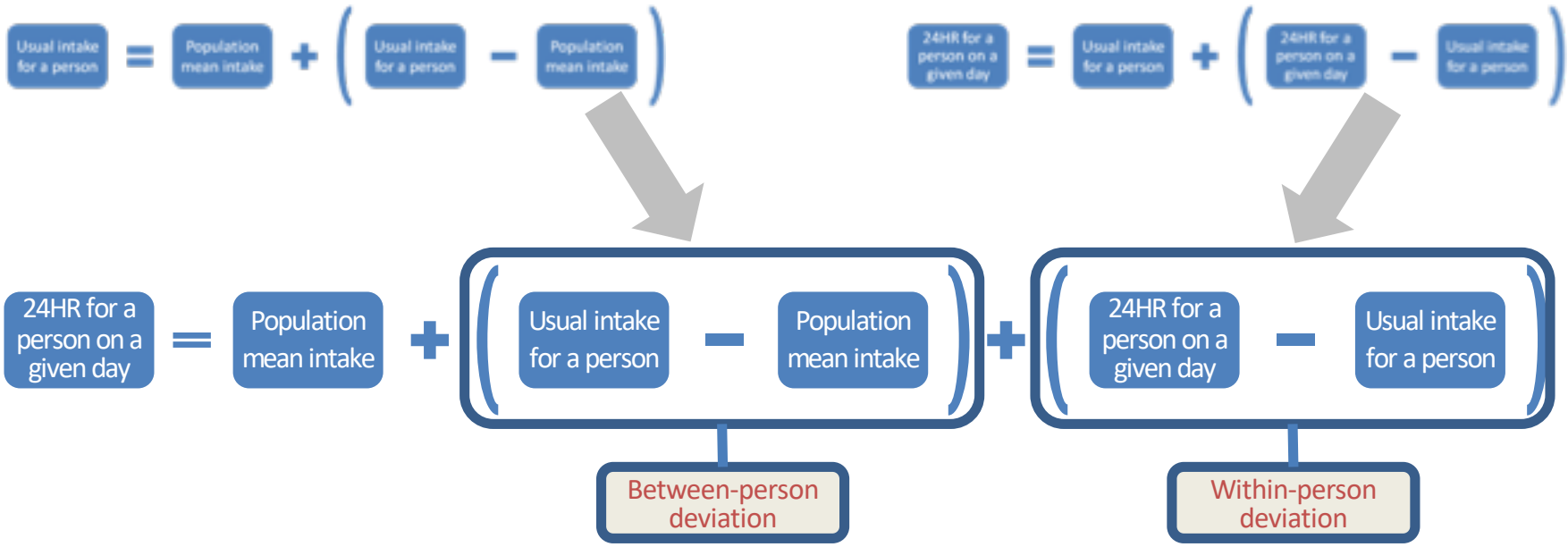
Foundation of the model

- Each usual intake is the population mean intake plus a deviation



- Between-person deviations cancel out across the population

Foundation of the model



- Population mean may depend on **fixed parameters**
- Both types of deviations are **random variables** with
  - Zero expectation
  - Unknown variances, distributions

Common variance assumption

- Sample variance among the 24HRs for a person estimates his within-person variance
  - Very few “degrees of freedom”, not very precise

- Assume same magnitude of within-person variation across individuals
  - Pool individual estimates to get more precise estimate

Distributional assumptions

- Statistically convenient to assume that both types of deviations follow a parametric probability distribution

- The normal distribution is a common choice
  - Only two parameters: mean and variance
  - Dependence between deviations can be completely modeled via correlation

Model for non-episodically consumed components

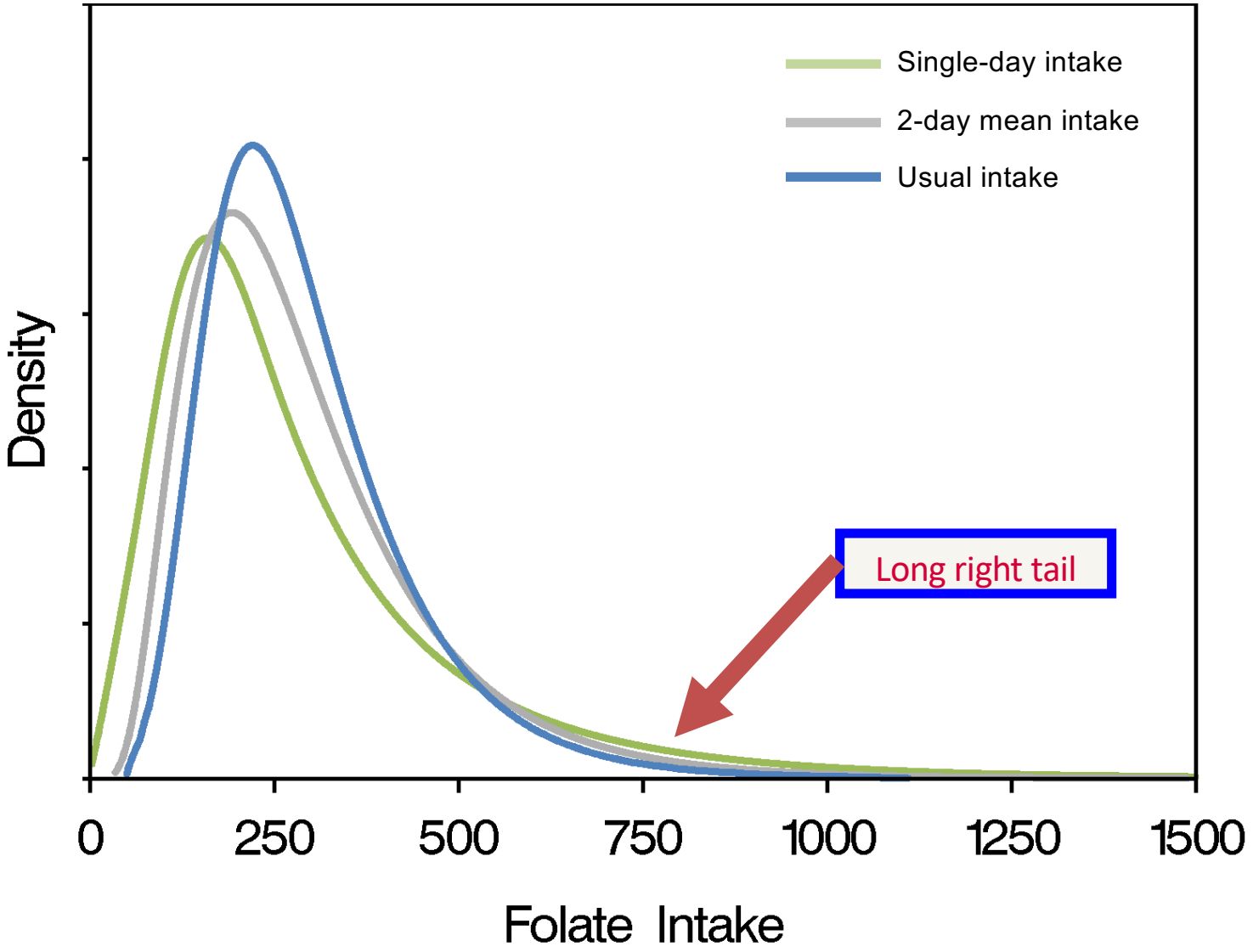
Within-person deviations are:

- Normally distributed, with a common variance
- Independent from those of other people
- Independent from those of the same person

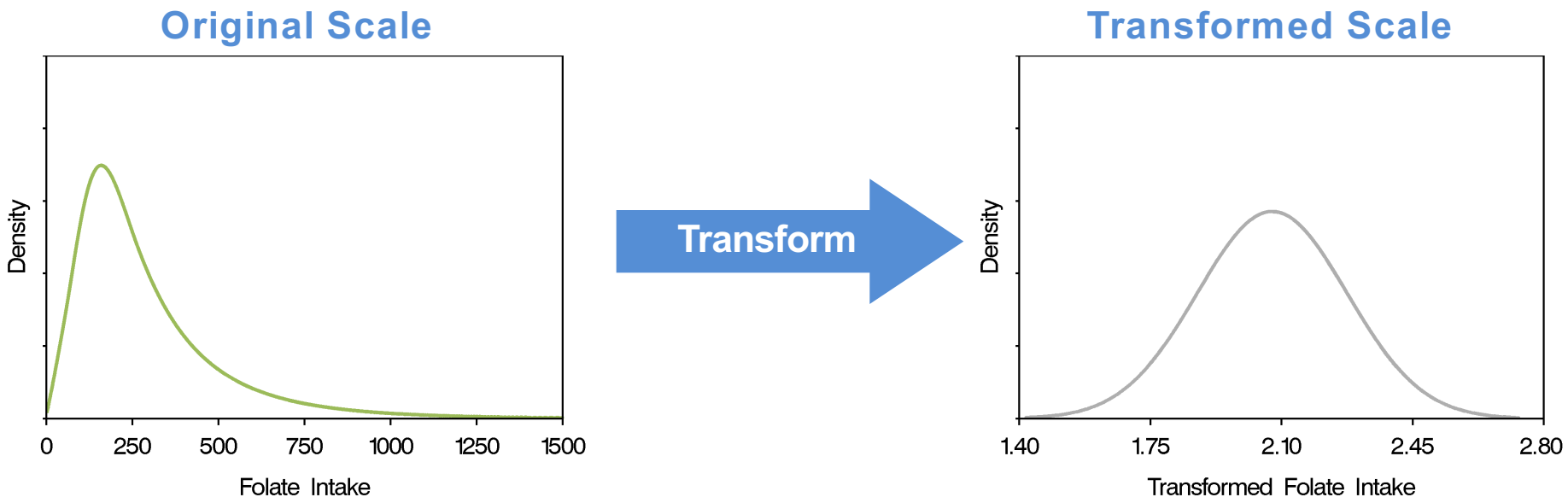
Between-person deviations are:

- Normally distributed, with a common variance
- Independent from those of other people

Distributions are often non-normal (skewed)



Nonlinear transformations to normality



- Power, Box-Cox, log, etc.
- Large values affected more than small ones
- Transformations must be invertible

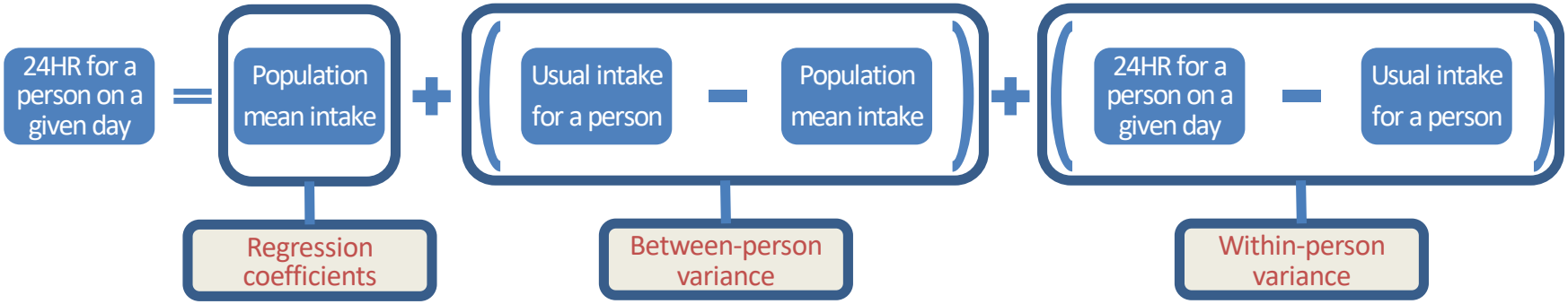
Features of the model

- Recognizes that unbiased ≠ error-free
- Corrects for within-person variation
- Requires additional assumptions
  - Common variance components
  - Normal within/between deviations
- Skewed distributions of intake may be handled with transformations

Limitations of the model

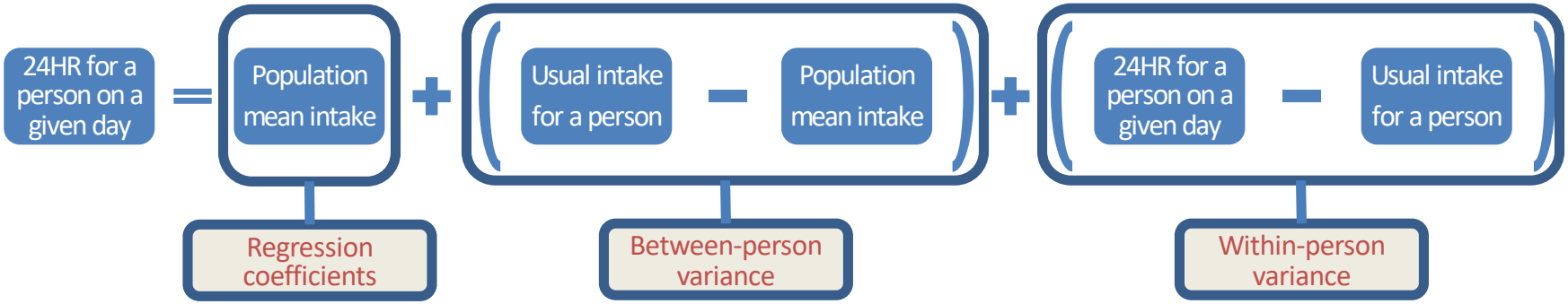
- Only estimates (conditional) means, variances of distributions for within/between-person deviations
- Model cannot estimate the two separate deviations for any particular person
- **No free lunch** – statistical model cannot provide a good estimate for a particular person's usual intake based on a handful of 24HRs

Estimating mean usual intake

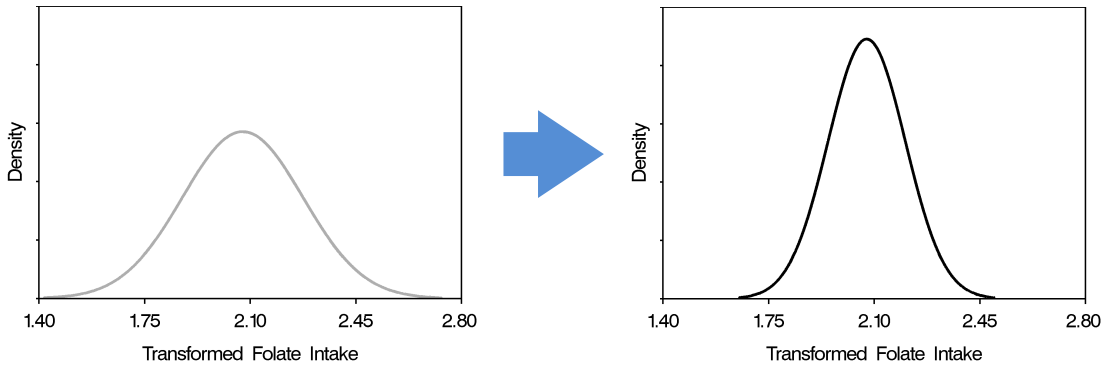


- Regression coefficients ("fixed effects") express mean 24HR in terms of covariates
- Unbiasedness assumption: mean 24HR is average usual intake
- Estimates the effect of covariates on usual intake

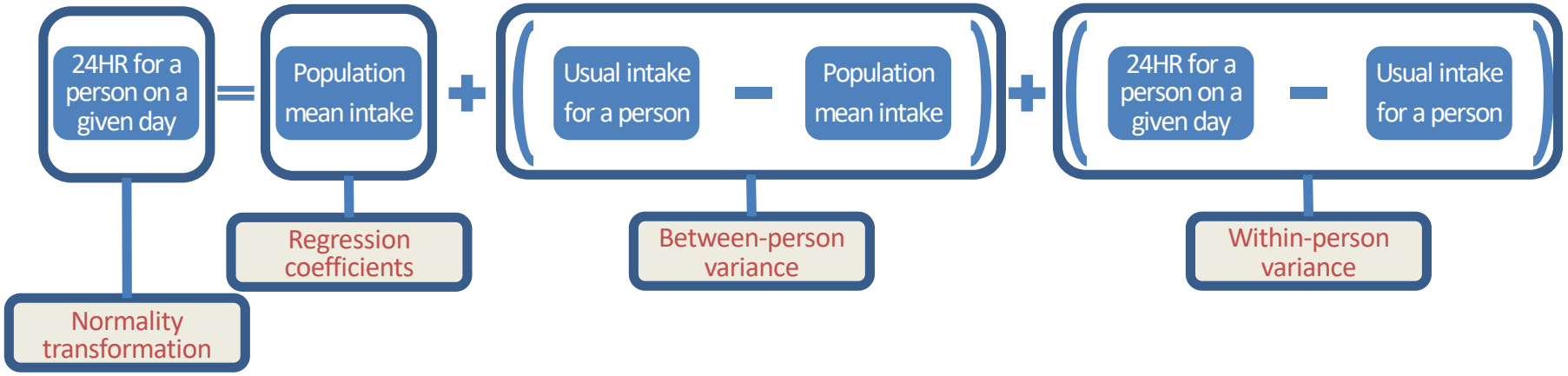
Estimating distributions of usual intake



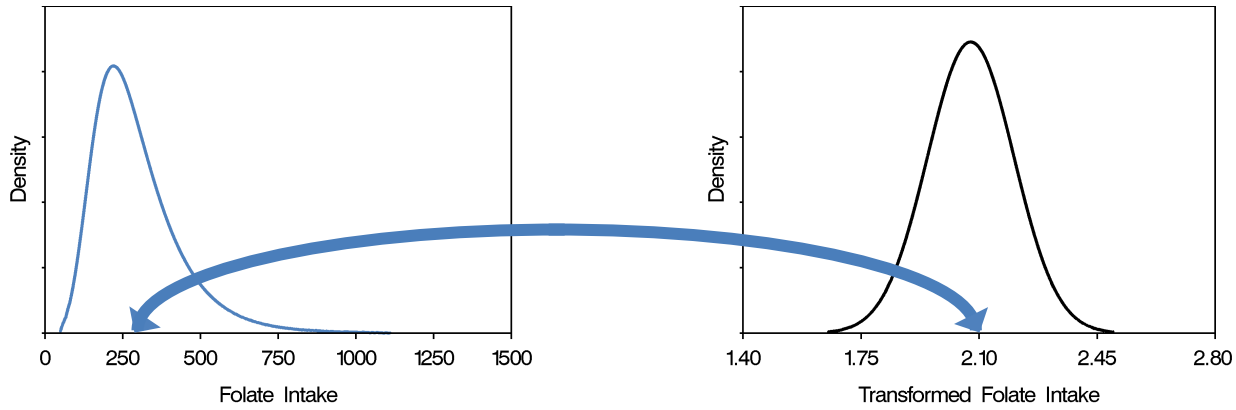
- Separates within- from between-person variation
- Can estimate usual intake distribution that exhibits only between-person variation



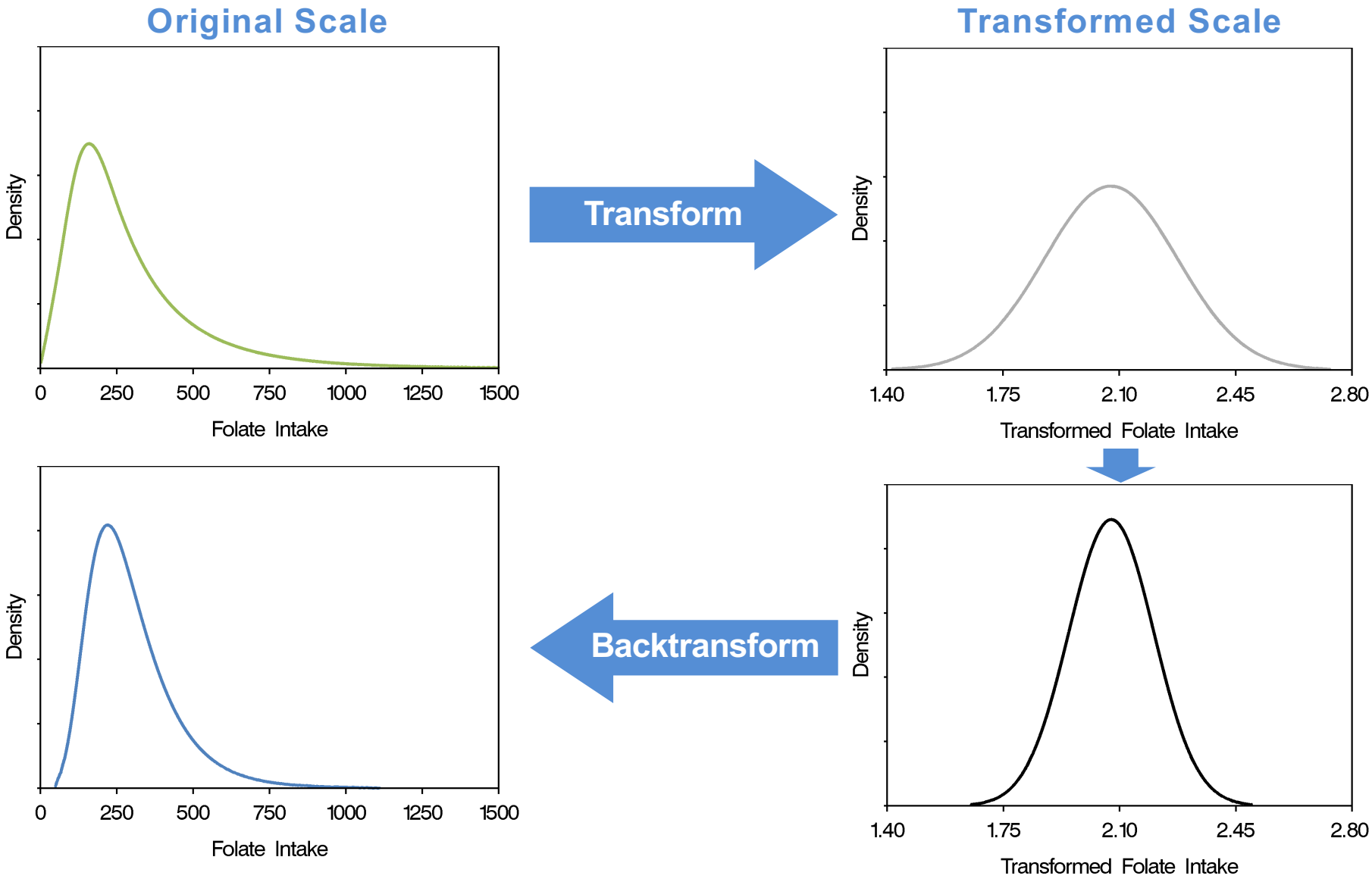
Estimating quantiles when transformations are used



- Estimate quantile of usual intake corresponding to one in the normal distribution that exhibits only between-person variance



Accounting for nonlinear transformations



MIXTRAN and DISTRIB

- MIXTRAN: estimates the mean (possibly as a function of covariates) and the variances of the within- and between-person deviations, *on a transformed scale*

```
%MIXTRAN(data=cchs, response=vitC,
foodtype=vitC, subject=seqn, repeat=day, outlib=out,
modeltype=amount);
```

- DISTRIB: uses the MIXTRAN output and simulates a population of true intakes, *in the original scale*

```
%DISTRIB (modeltype=amount,
pred=out._pred_vitC, param=out._param_vitC,
outlib=out, subject=seqn, food=vitC);
```

Regression calibration for diet-health models

- Regression using error-prone exposure measurements (i.e., average of two 24HR) usually produces slope estimates biased toward the null – “attenuated”

- Statistical model for 24HRs can be used to construct “calibrated” measurements that can be used in place of error-prone measurements to provide an unbiased slope estimate (INDIVINT)

- Often called (**erroneously**) “individual usual intakes” (**OOPS!**)

Non-episodic versus episodic

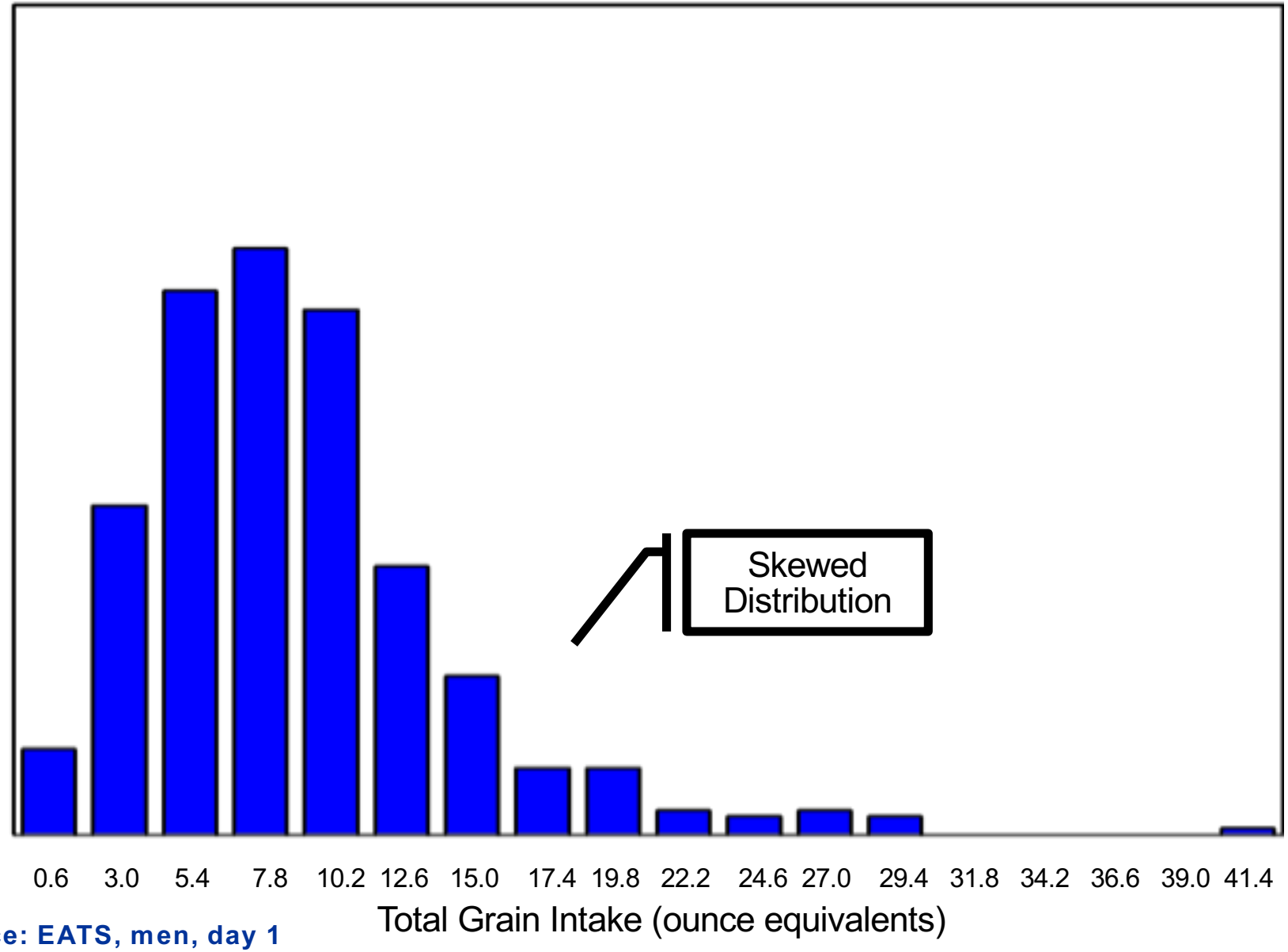
- What makes up your diet
  - Foods
  - Food groups

Many are "episodically consumed"

  - Components of foods
    - Macronutrients
    - Micronutrients

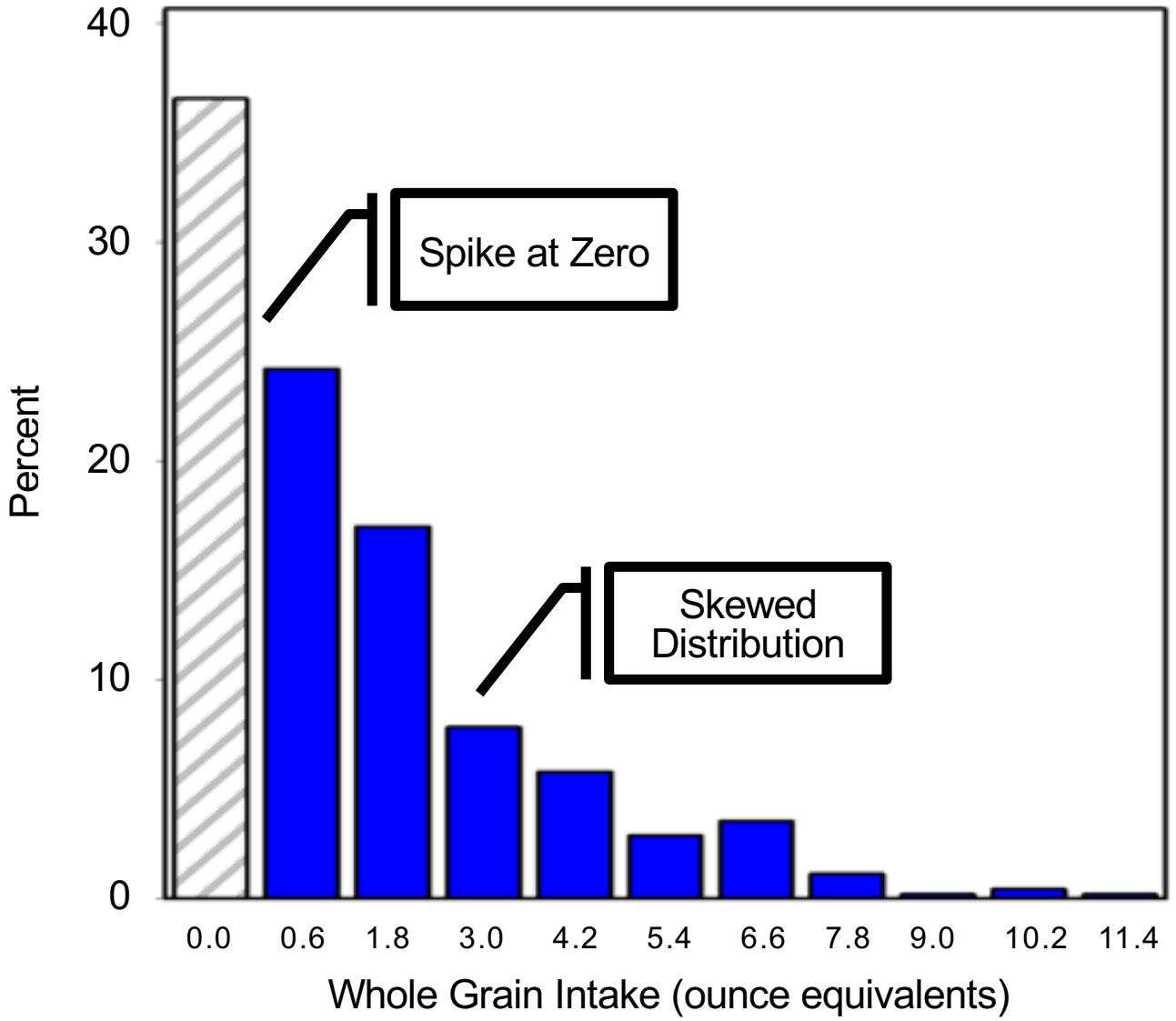
Most are consumed daily by most persons

Some foods are consumed every day…

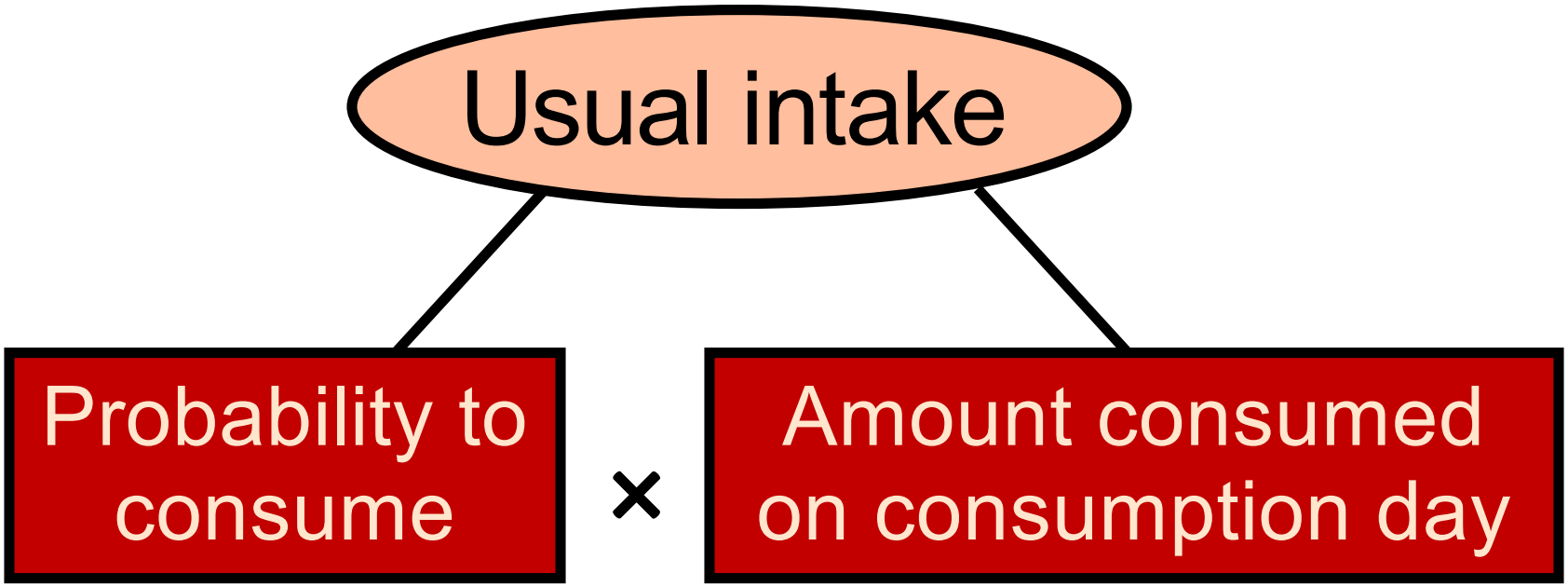


Source: EATS, men, day 1

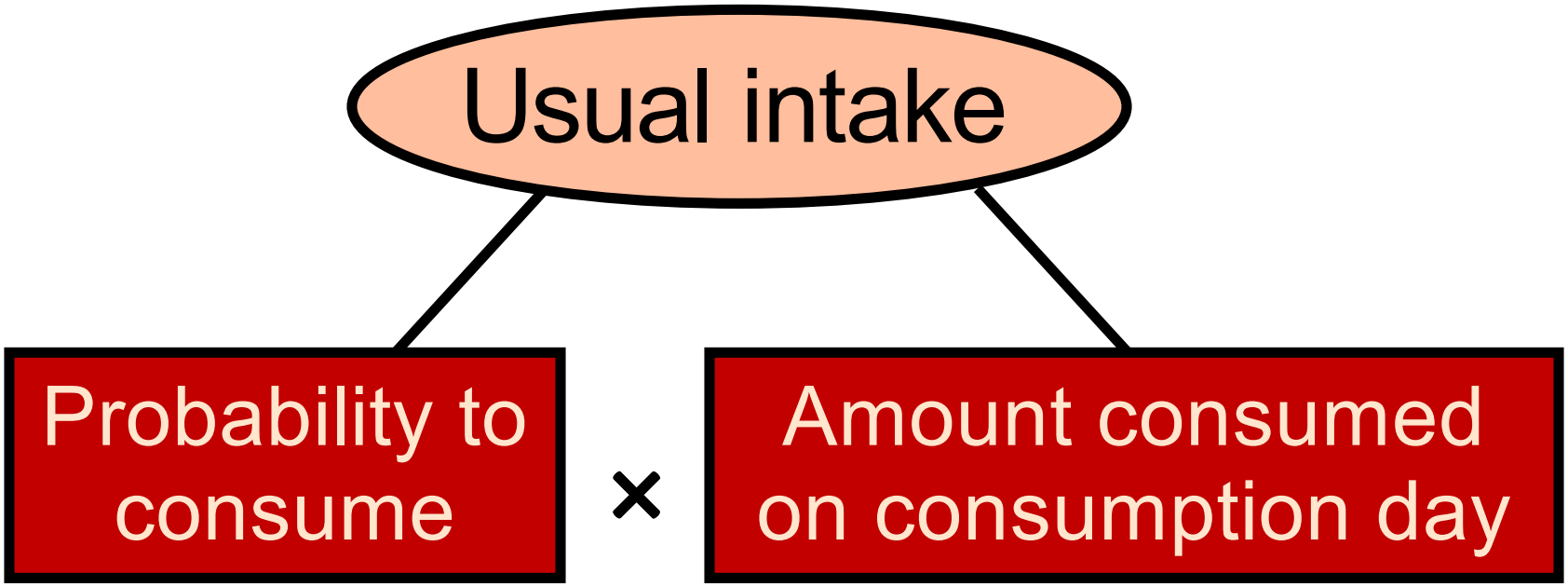
But many other foods are consumed episodically



Two parts of usual intake



Statistical models needed for each part

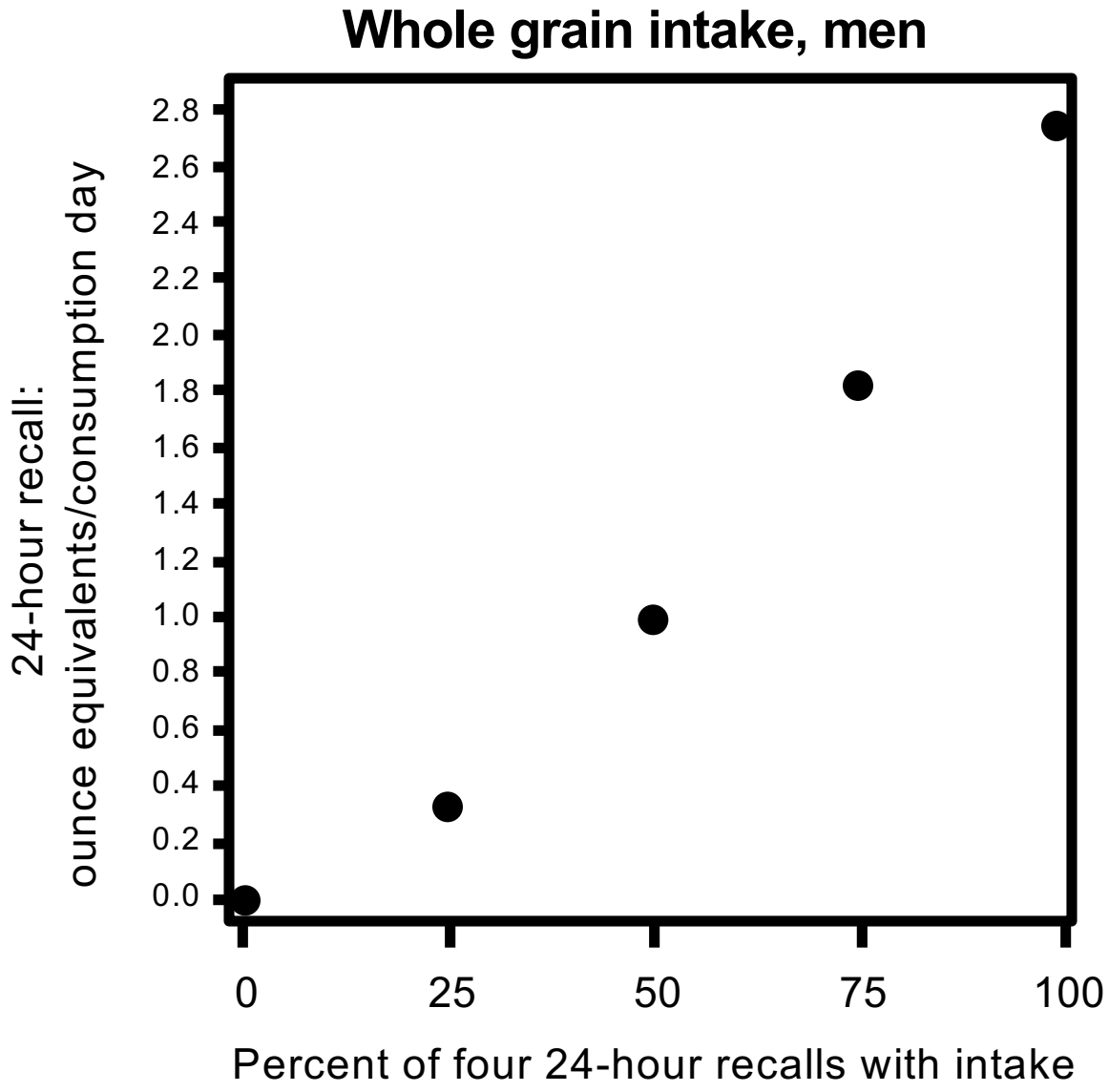


- Amounts consumed on consumption days follow the basic or one-part model, with group-level means and both between- and within-person deviations

Model for probability to consume

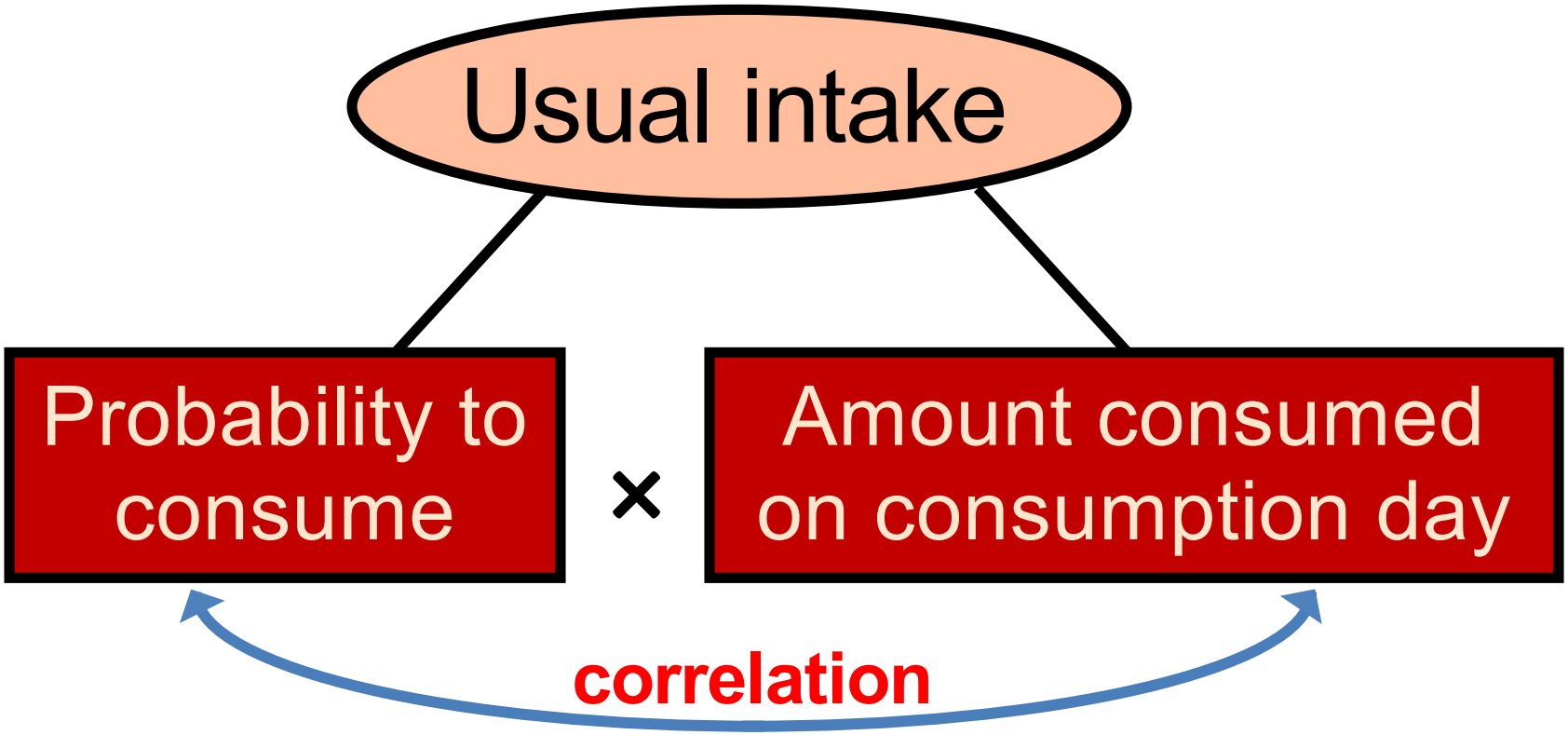
- Event of consumption on a day is binary, determined by a “coin flip”
- Group-level probability to consume may be influenced by covariates
- A person’s probability to consume is a combination of group-level probability and a **between-person deviation**
- Modeled via logistic regression

Correlation between probability and amount



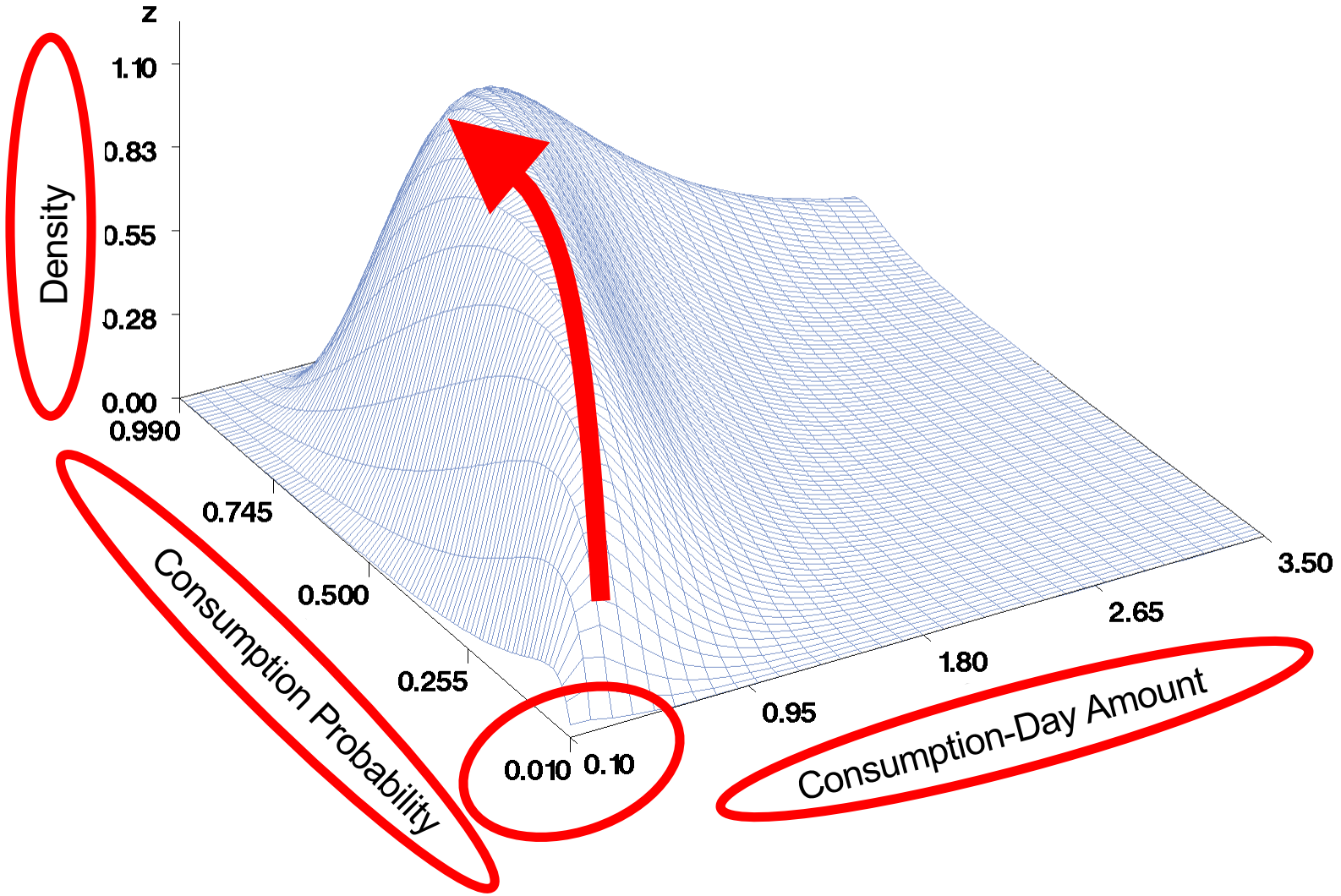
Source: EATS

Inducing correlation between the parts of the model



- Models for each part may share covariates
- Between-person deviations in the two models allowed to be correlated

Joint distribution of probability and amount



Using the two-part model

- Fixed effect equations estimate how probability to consume and amount consumed are related to their respective sets of covariates
- Joint distribution of probability and amount provides marginal distribution of either quantity, and also functions thereof (e.g. their product)
- Model can also be used to derive calibrated measurements for use in diet/disease models

MIXTRAN and DISTRIB – One-part model

- MIXTRAN: estimates the mean (possibly as a function of covariates) and the variances of the within- and between-person deviations, *on a transformed scale*

```
%MIXTRAN(data=cchs, response=vitC,
foodtype=vitC, subject=seqn, repeat=day, outlib=out,
modeltype=amount);
```

- DISTRIB: uses the MIXTRAN output and simulates a population of true intakes, *in the original scale*

```
%DISTRIB (modeltype=amount,
pred=out._pred_vitC, param=out._param_vitC,
outlib=out, subject=seqn, food=vitC);
```

MIXTRAN and DISTRIB – Two-part model

- MIXTRAN: estimates the mean (possibly as a function of covariates) and the variances of the within- and between-person deviations, *on a transformed scale*

```
%MIXTRAN(data=cchs, response=wholegrains,
foodtype=wholegrains, subject=seqn, repeat=day,
outlib=out, modeltype=corr);
```

- DISTRIB: uses the MIXTRAN output and simulates a population of true intakes, *in the original scale*

```
%DISTRIB (modeltype=corr,
pred=out._pred_wholegrains,
param=out._param_wholegrains, outlib=out,
subject=seqn, food=wholegrains);
```

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics

Dietary Reference Intakes: Estimated Average Requirements

Food and Nutrition Board, Institute of Medicine, National Academies

Life Stage Group	Calcium (mg.d)	CHO (g/d)	Protein (g/kg/d)	Vit A (µg/d)	Vit C (mg/d)	Vit D (µg/d)	Vit E (mg/d)	Thiamin (mg/d)	Riboflavin (mg/d)
Infants									
0 - 6 mo									
6 - 12 mo			1.0						
Children									
1-3 y	500	100	0.87	210	13	10	5	0.4	0.4
4-8 y	800	100	0.76	275	22	10	6	0.5	0.5
Males									
9-13 y	1,100	100	0.76	445	39	10	9	0.7	0.8
14-18 y	1,100	100	0.73	630	63	10	12	1.0	1.1
19-30 y	800	100	0.66	625	75	10	12	1.0	1.1
31-50 y	800	100	0.66	625	75	10	12	1.0	1.1
51-70 y	800	100	0.66	625	75	10	12	1.0	1.1
> 70 y	1,000	100	0.66	625	75	10	12	1.0	1.1
Females									
9-13 y	1,100	100	0.76	420	39	10	9	0.7	0.8
14-18 y	1,100	100	0.71	485	56	10	12	0.9	0.9
19-30 y	800	100	0.66	500	60	10	12	0.9	0.9
31-50 y	800	100	0.66	500	60	10	12	0.9	0.9
51-70 y	1,000	100	0.66	500	60	10	12	0.9	0.9
> 70 y	1,000	100	0.66	500	60	10	12	0.9	0.9

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics
- Population monitoring:
  - Characterizing *a priori* "at-risk" subpopulations
  - Proportion not meeting sex/age-specific targets vs. not meeting "average" target

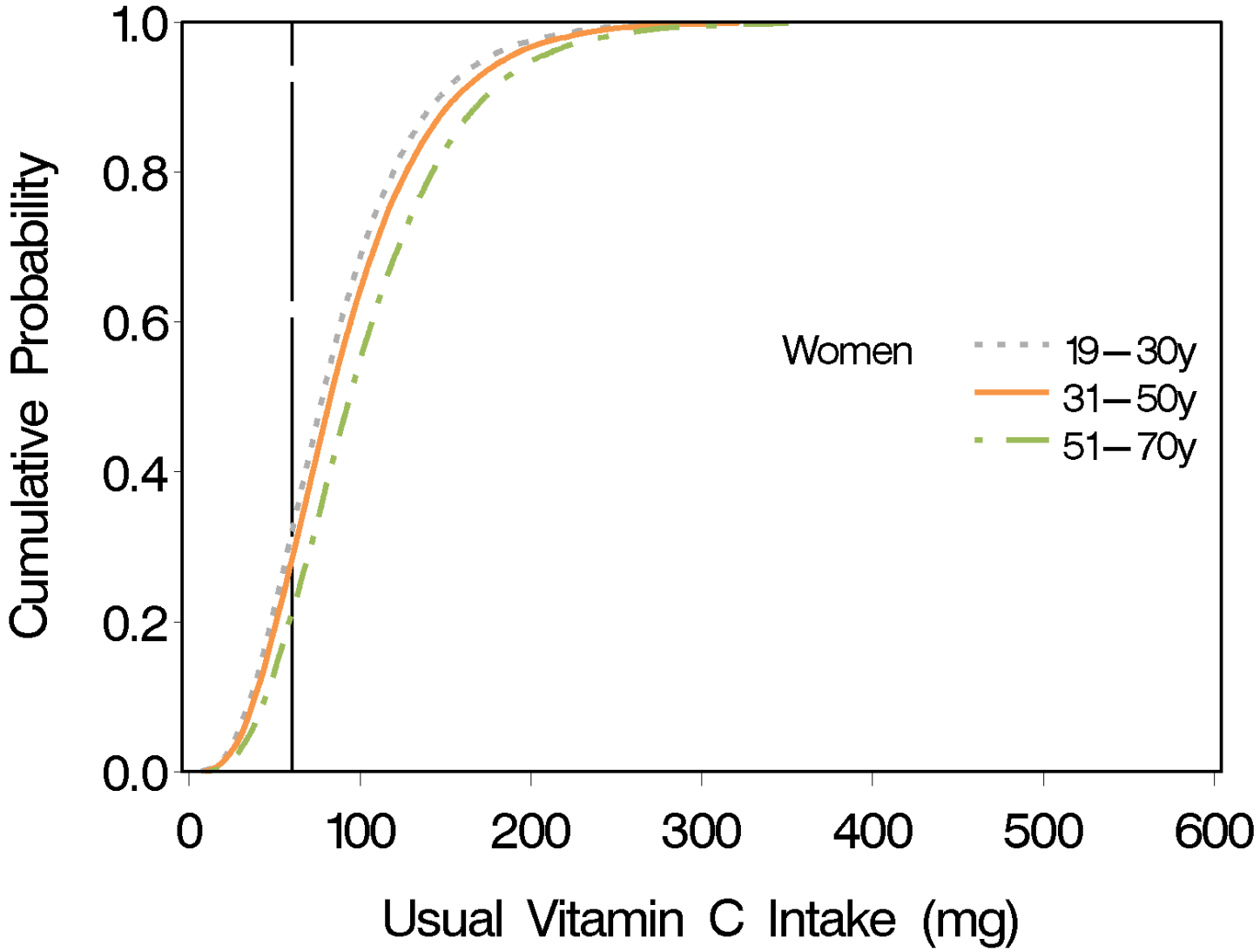
One solution is stratification

- Run separate analyses on subsamples defined by personal characteristics
  - Population proportion not meeting sex/age targets is weighted average of subpopulation proportions
- Small subsamples lead to less precise estimates

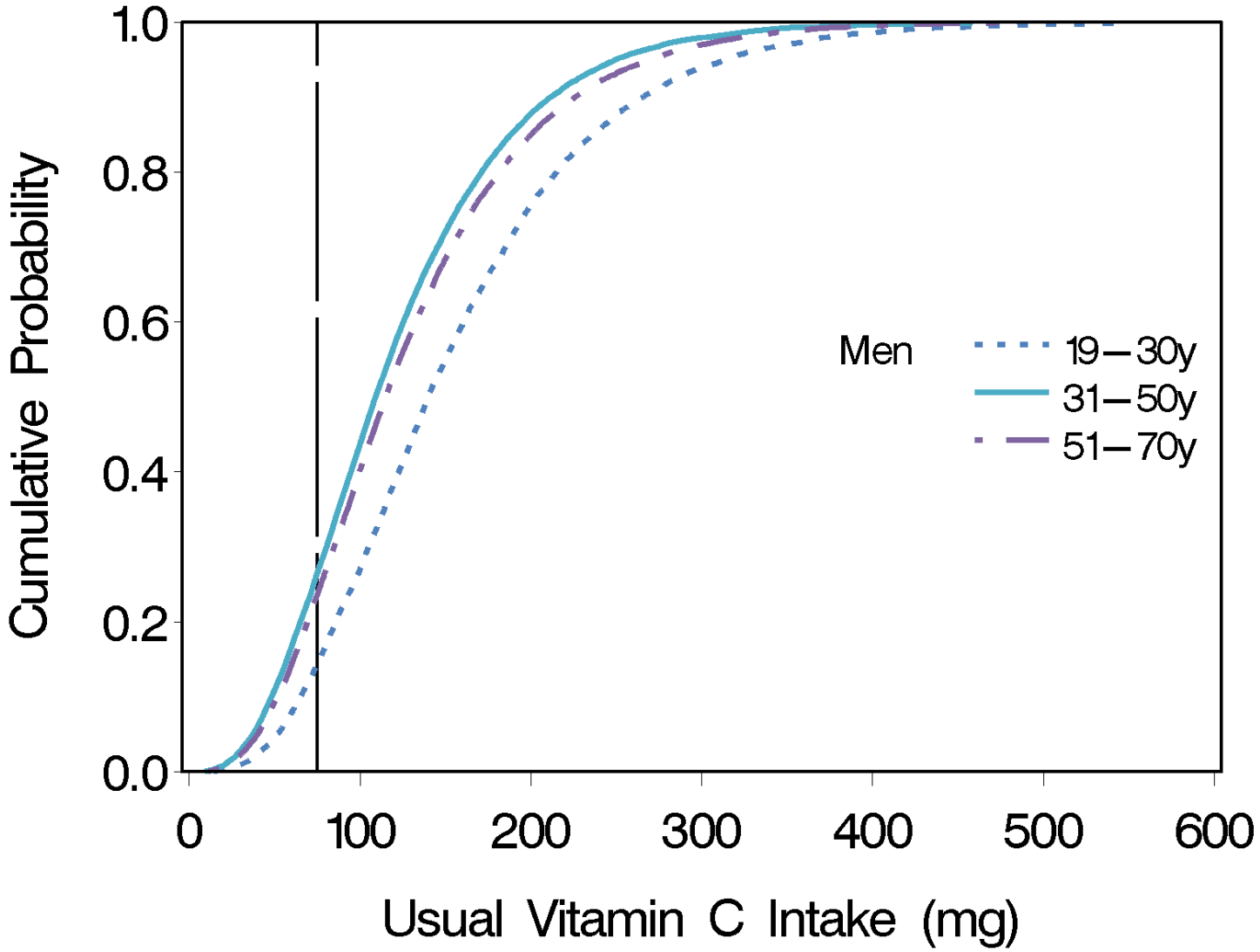
The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics
- Population monitoring:
  - Characterizing *a priori* "at-risk" subpopulations
  - Proportion not meeting sex/age-specific targets vs. not meeting "average" target
- Understanding determinants of diet
  - Identify characteristics associated with higher/lower average intake, e.g., smoking

Example: Eating at America's Table



Example: Eating at America's Table



MIXTRAN and DISTRIB – With covariates

- MIXTRAN: estimates the mean (possibly as a function of covariates) and the variances of the within- and between-person deviations, *on a transformed scale*

```
%MIXTRAN(data=cchs_women, response=vitC,
foodtype=vitC, subject=seqn, repeat=day,
outlib=out_women, covars_amt = age1 age2;
modeltype=amount);
```

- DISTRIB: uses the MIXTRAN output and simulates a population of true intakes, *in the original scale*

```
%DISTRIB(modeltype=amount,
pred=out_women._pred_vitC,
param=out_women._param_vitC, outlib=out_women,
subject=seqn, food=vitC);
```

Limitations of stratification approach

- When multiple factors thought to influence diet are considered,
  - Subsample sizes decrease dramatically
  - Analysis burden increases

- Allowing covariates in the statistical models can overcome this limitation

Types of covariates

- **Individual-level**: affects true intake on all days, e.g., gender, age, smoker/nonsmoker status
- **Time-dependent**: affects true intake on specific days, e.g., season, weekday
- **Nuisance**: affects reporting error, e.g., interview sequence, mode (telephone vs. in-person)

Agenda and logistics

- **Break**
- Estimation of distributions of usual intake, cont'd – *Kevin Dodd*
- Q&A
- Early experience applying the NCI method to CCHS 2015 – *Karelyn Davis* and *Cunye Qiao*
- Reflections from *Kevin Dodd*
- Q&A
- Afternoon: in-person or webcast

Elements of complex sampling designs

- Stratification
- Clustering
- Weighting

Weighting for multi-stage samples

- Each individual's **sample weight** is equal to the inverse of the final probability of being selected from the population

$$\text{sample weight} = \frac{1}{\text{final probability}}$$

$$\begin{aligned}\text{final probability} = \;& \text{probability of county being selected} \\ & \times \text{probability of segment being selected from county} \\ & \times \text{probability of household being selected from segment} \\ & \times \text{probability of individual being selected from household}\end{aligned}$$

Additional considerations for weighting

- Can incorporate
  - Differential selection probabilities due to stratification and clustering
  - Differential nonresponse probabilities

- Weighted counts of sampled individuals with particular demographic characteristics often set to reproduce "known" population counts – **poststratification**

All survey design elements must be accounted for

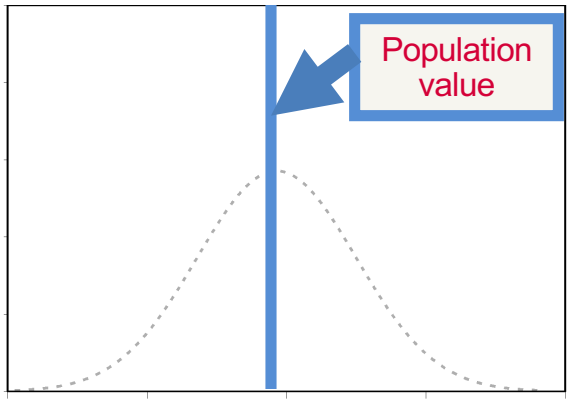
- Weighting required to minimize bias in survey-based population estimates
- Stratification, clustering, and weighting affect standard errors of estimates

Statistical methods for complex surveys limited

- Inference based on t-tests easiest to extend to complex surveys
  - Asymptotic normality, standard error formulae established for many mean-like statistics
- Other statistical methods more difficult to extend
  - E.g., likelihood ratio tests

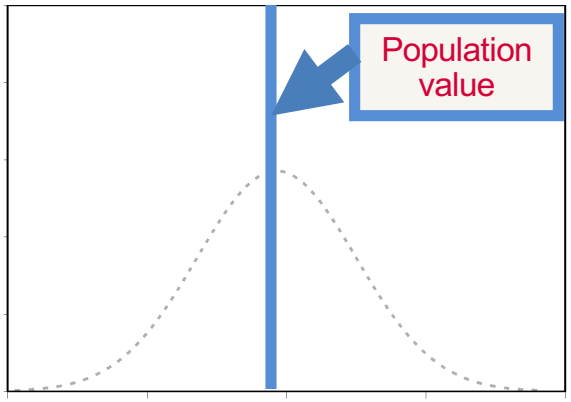
What is the standard error of an estimate?

- Reflects variation expected across repeated sampling of the population
  - Most samples yield estimates close to true population value, a few samples yield estimates far away
  - Sampling distributions are often normal (CLT)



What is the standard error of an estimate?

- The **standard error** (s.e.) is the standard deviation of the sampling distribution
  - More independent pieces of information ⇒ smaller standard errors
- Used to construct significance tests, confidence intervals assuming **asymptotic** normality



Estimating standard errors

- In practice, only one sample is obtained
  - Standard errors must be estimated from the data at hand

- Basic statistical theory provides estimation methods for standard errors of “smooth” statistics
  - Means
  - “Mean-like”: regression parameters, ratios

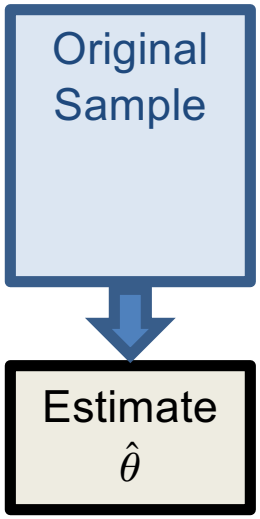
Estimating standard errors

- Estimating standard errors for percentiles is especially challenging
  - Not “mean-like” for purposes of CLT
  - Sampling distributions less well-behaved

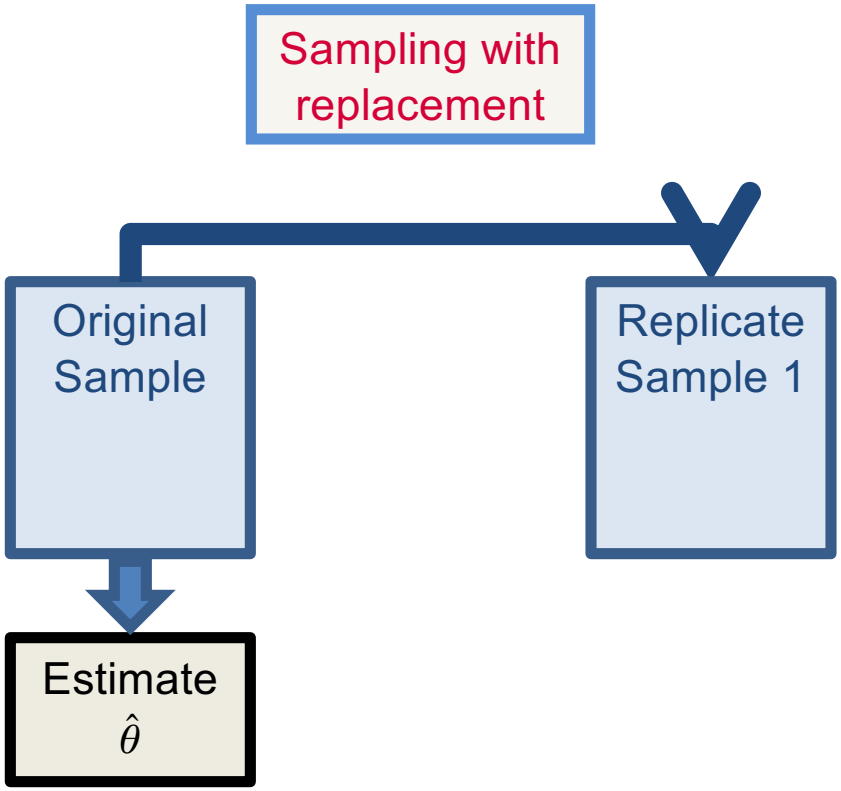
- May require alternative methods for tests/CIs
  - Standard error still reflects variation over repeated sampling

Example: Bootstrap in simple random sampling

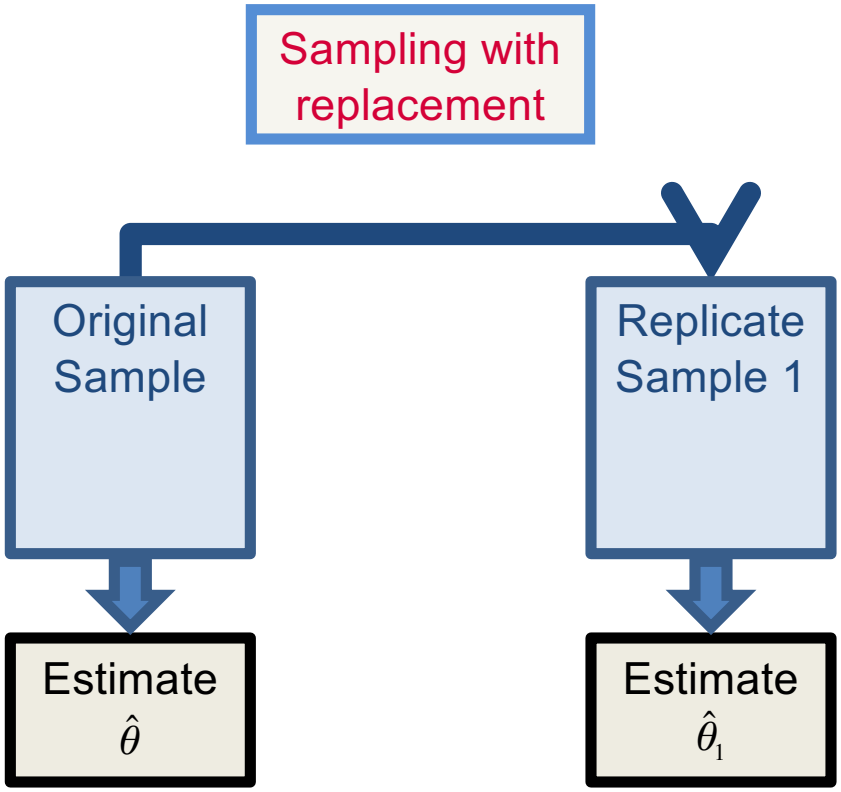
Example: Bootstrap in simple random sampling



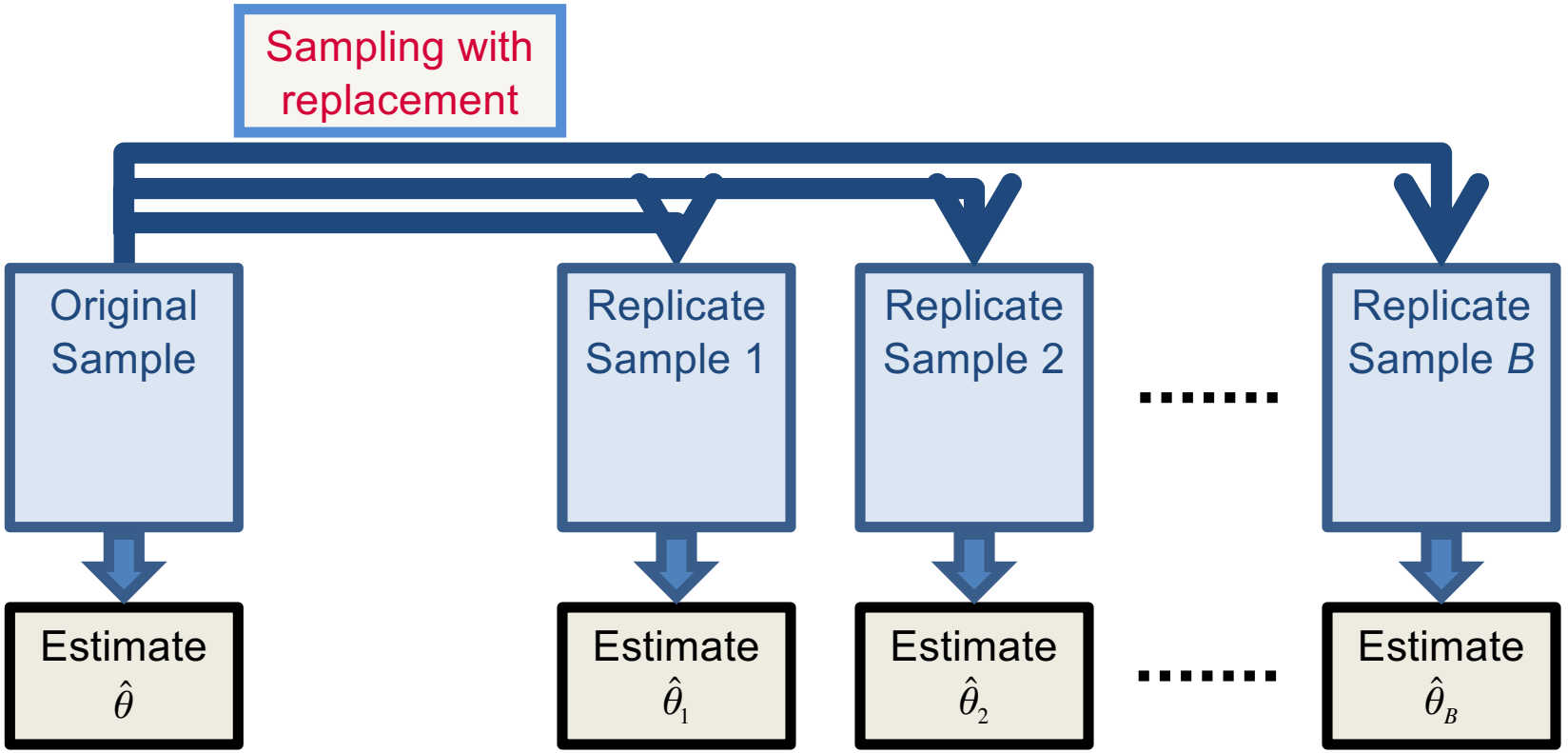
Example: Bootstrap in simple random sampling



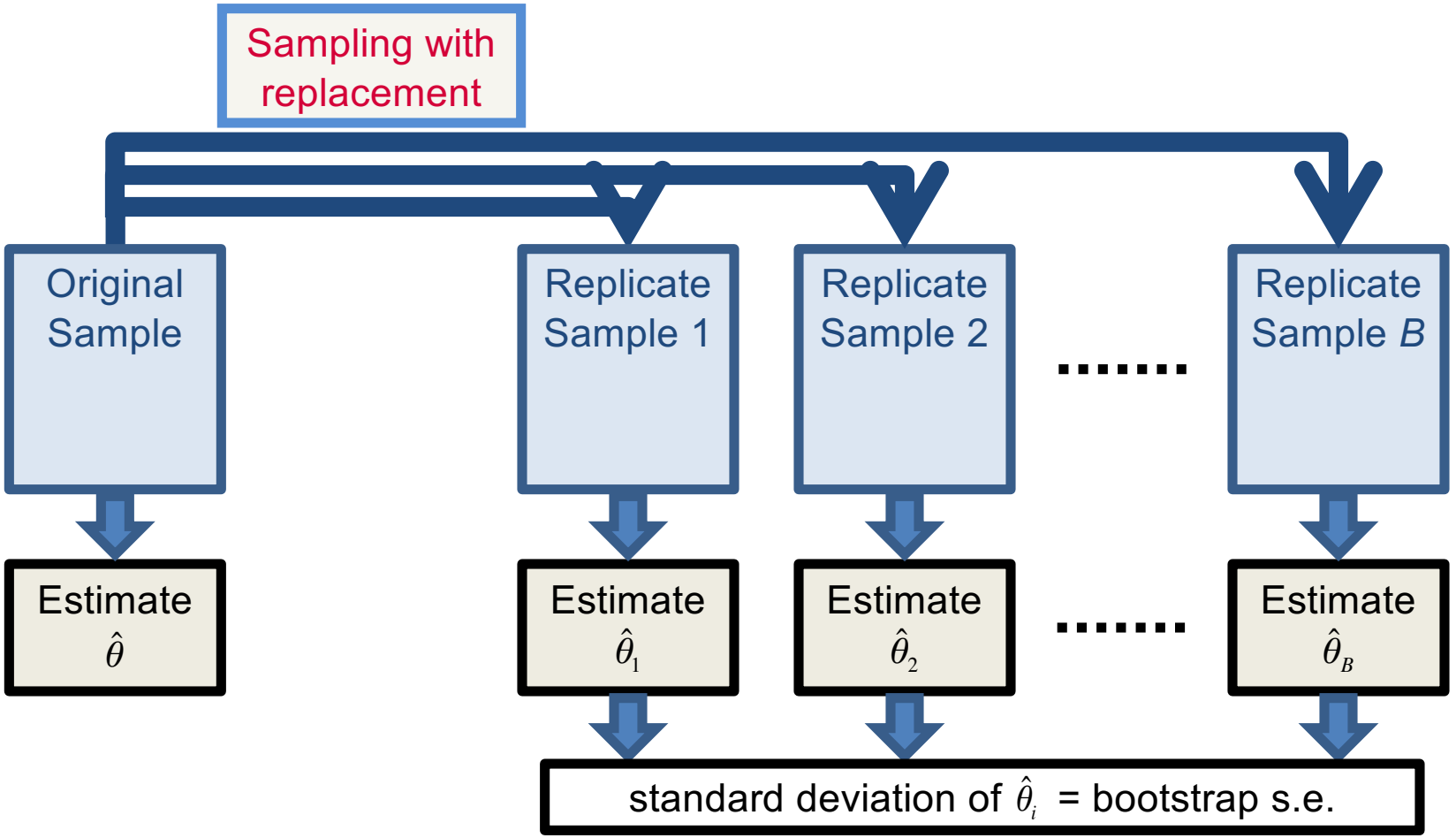
Example: Bootstrap in simple random sampling



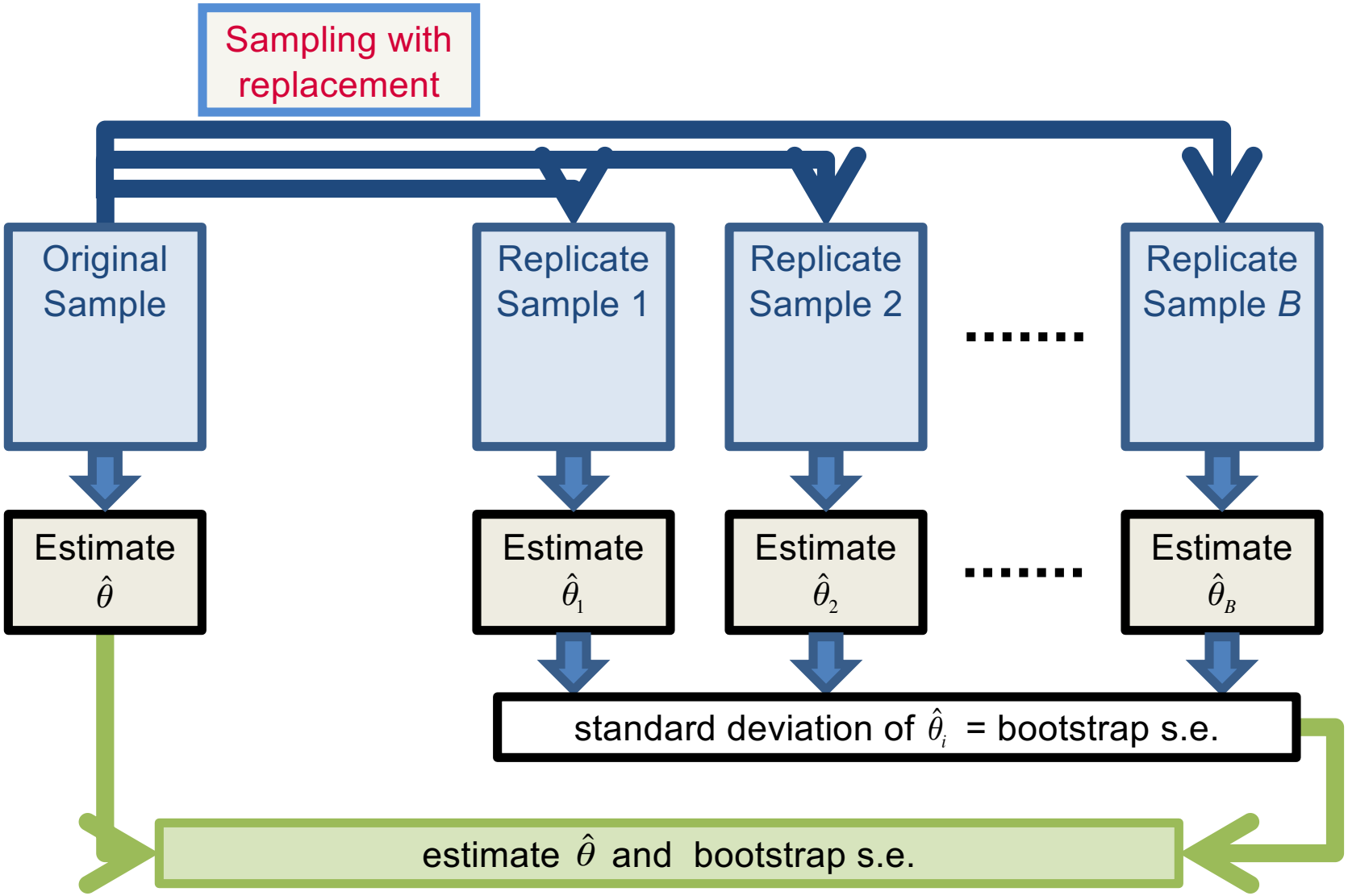
Example: Bootstrap in simple random sampling



Example: Bootstrap in simple random sampling



Example: Bootstrap in simple random sampling



Example: Bootstrap in simple random sample

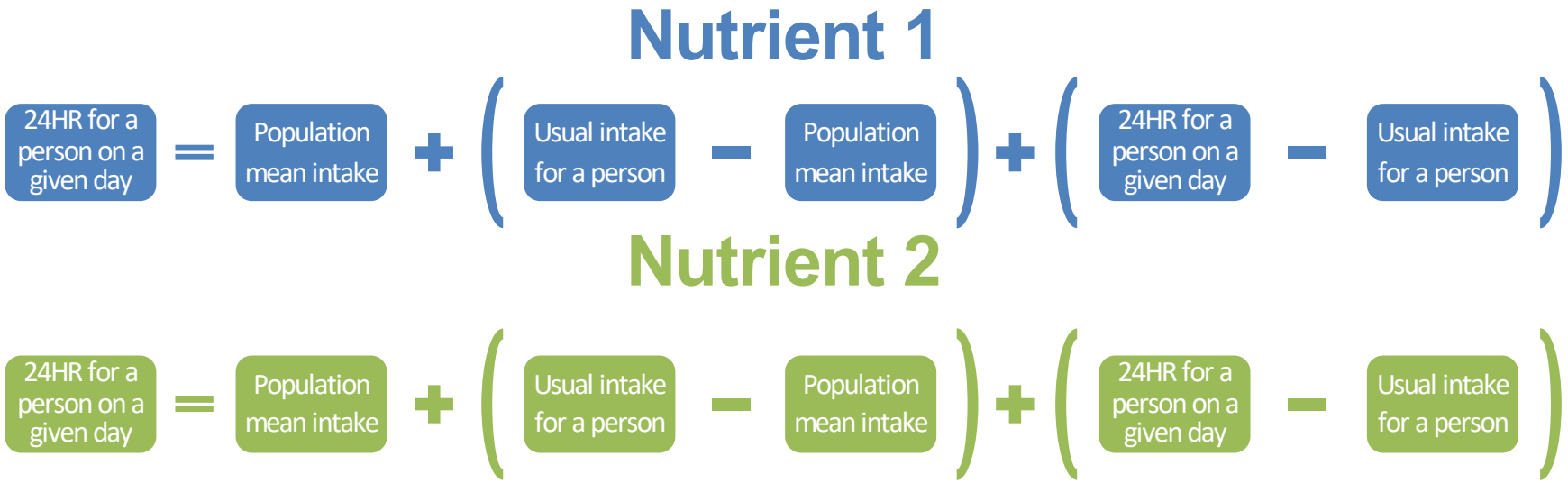
- Key to bootstrap is **with-replacement sampling**
- In a given bootstrap sample,
  - Some individuals will appear multiple times
  - Some individuals will not appear at all
- Number of times an individual appears is analogous to a sampling weight

Bootstrap in complex surveys

- Bootstrap samples must be drawn according to sampling plan used to draw the original sample
  - Computationally intensive (*B* very large, i.e., 500 bootstraps for CCHS)

- Offers robust method for constructing CIs
  - Bounds based on 95% of empirical distribution of bootstrap estimates
  - May work better for poorly-behaved sampling distributions of “non-smooth” statistics

Extensions to handle multiple components



- Joint models for multiple dietary components must allow for correlation
  - among within-person deviations, because they are based on data from the same 24HR
  - among between-person deviations, because they are based on data from the same person

Extensions to handle multiple components

- Model generalizes to handle combinations of episodic and non-episodic dietary components
- Logistic model is replaced with probit model to allow correlation between event of food consumption and amount of nutrient consumed on the same day
- Model can be used to estimate multivariate distributions and calibrate multiple dietary variables in diet/health models – **dietary patterns**

Summary

- Statistical methods may be used to model usual intake using repeated applications of short-term instruments such as 24HRs
  - Estimation is a measurement error problem

- Methods rely on assumed unbiasedness of short-term instruments
  - If the assumption does not hold, estimates may still be biased,
  - But are likely to be closer to the truth than if measurement error was ignored

NATIONAL CANCER INSTITUTE

www.cancer.gov

www.cancer.gov/espanol

Agenda and logistics

- **Q&A**
- Early experience applying the NCI method to CCHS 2015 – *Karelyn Davis* and *Cunye Qiao*
- Reflections from *Kevin Dodd*
- Q&A
- Afternoon: in-person or webcast