Estimating usual intake distributions for dietary components consumed daily by nearly all persons

Kevin W. Dodd, PhD
National Cancer Institute

Objectives

Two main areas of interest

- Describing usual intake distributions: mean, percentiles, proportion above or below a threshold
- Estimating diet-health relationships: regression coefficients

Participants will have an understanding of:

- Considerations in estimating usual intakes of nutrients and foods consumed nearly daily by nearly all persons
- Assumptions made in current approaches to estimating usual intake distributions
- Statistical modeling techniques and data requirements for estimating usual intake distributions

Webinars 6-8, 12
Objectives

**Daily versus episodic consumption**

- Consumed nearly daily by nearly all persons
  - E.g., vitamin C, total grains, total vegetables, solid fats, added sugars
- Consumed episodically by most persons
  - E.g., vitamin A, whole grains, dark green vegetables, fish

**Objectives**

- Daily versus episodic consumption
- Estimating usual intake distributions for dietary components consumed daily by nearly all persons
  - E.g., vitamin A, whole grains, dark green vegetables, fish

Outline

- Basic assumptions
- Building a statistical model
- Estimating distributions from the model
- The role of covariates

**BASIC ASSUMPTIONS**

- Focus is on usual intake
  - Usual intake = long-term average daily intake
  - Reflects idea that nutritional goals should be met over time, but not necessarily every day
  - Provides a measure of total (chronic) exposure
    - Not addressing issues of acute exposure here

**Basic assumptions**

- Challenge
  - Usual intakes are not directly observable
    - Self-report dietary assessment instruments measure usual intake with error
    - If ignored, this error can bias results
    - Statistical modeling methods can be used to correct this bias
Assessment strategies fall between two extremes

Usual intake = long-term average daily intake
- Focus on long-term aspect
  - Food Frequency Questionnaire (FFQ)
- Focus on daily aspect
  - 24-hour recall (24HR)

Two classes of measurement error in instruments
- Random: Average of repeats = true value
  - Instrument is accurate, or unbiased
  - May not be precise
- Systematic: Average of repeats ≠ true value
  - Instrument is inaccurate, or biased
  - Systematic bias can occur in many ways

Potential sources of error in instruments: FFQ
- Cognitively challenging
- Limited food list/portion size choices
- No need for repeated application
  (high reproducibility)

Potential sources of error in instruments: 24HR
- Less cognitively challenging
- Open-ended format
- Repeats required to deal with day-to-day variation in intake

Comparison of measurement error structures

<table>
<thead>
<tr>
<th>24-hour recall (24HR)</th>
<th>Food frequency questionnaire (FFQ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Larger within-person random error</td>
<td>• Smaller within-person random error</td>
</tr>
<tr>
<td>• Smaller systematic error</td>
<td>• Larger systematic error</td>
</tr>
</tbody>
</table>

Rationale for using short-term instruments
- Effects of random error can be mitigated by averaging repeats
  - Modeling can perform the same task
- Effects of systematic error cannot be mitigated unless we have an additional instrument
- Therefore, usual intake distributions based on 24HRs should be closer to those of truth than those based on FFQ
Basic assumptions

**Main assumption**

24HR unbiased for individual-level usual intake

![Graph showing true usual intake, average reported intake, and days](image)

**Working assumption**

- Unbiasedness of 24HR is a *working assumption*
- Required to proceed with development of methods
- May be more or less justified depending on dietary component of interest

**Building a model**

Typical data scenario

A small number of replicated 24HRs collected on each of many individuals

**Notation**

- Observations denoted by $R_{ij}$, usual intake by $T_i$
- Individuals indexed by subscript $i = 1, 2, \ldots, N$
- Repeats indexed by subscript $j = 1, 2, \ldots, J$

**Estimating usual intake distributions for dietary components consumed daily by nearly all persons**

**Implications of unbiasedness assumption**

24HR unbiased for individual-level usual intake

$$T_i = \text{E}[R_{ij} | i]$$

![Graph showing true usual intake, average reported intake, and days](image)

**Implications of unbiasedness assumption**

- The mean usual intake for the population is another kind of average:

$$\mu = \text{E}[T_i] = \text{E}[\text{E}[R_{ij} | i]]$$

- The population mean usual intake can be estimated as the average of within-person average 24HRs
Building a statistical model for 24HRs

Within- and between-person variation

- Effect of within-person variation

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

Limitations of unbiasedness assumption

- What about characteristics of the usual intake distribution other than the mean?
  - With only a few repeats, averaging only removes some of the within-person variation
  - Distributions of averages are too wide relative to usual intake distributions

Effect of within-person variation

- Population mean usual intake may be well estimated by simple averaging methods
- Percent of population with usual intake below/above cutoff values may be very biased – modeling necessary

Foundation of the model

- Each recall is usual intake plus a deviation

\[ R_y = T_i + (R_y - T_i) = T_i + e_y \]
Each recall is usual intake plus a deviation

\[ R_y = T_i + (R_y - T_i) = T_i + \epsilon_y \]

- **Within-person deviation**

Each usual intake is the population mean intake plus a deviation

\[ T_i = \mu + (T_i - \mu) = \mu + u_i \]

- **Between-person deviation**

Each usual intake is the population mean intake plus a deviation

\[ T_i = \mu + (T_i - \mu) = \mu + u_i \]

- Population mean \( \mu \) is a **fixed parameter**
- 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
  - Naturally parameterized by mean and variance
  - Dependence between deviations can be completely modeled via correlation

Basic statistical model for 24HRs

Within-person deviations are:

- Normally distributed, with a common variance
  - Can be relaxed, if desired

- Independent from those of other people

- Independent from those of the same person
  - Can be relaxed, e.g., if 24HRs are consecutive

Between-person deviations are:

- Normally distributed, with a common variance
  - Can be relaxed, if desired

- Independent from those of other people

Complications of skewed distributions

- Large values affected more than small ones
- Other transformations possible
  - Should be one-to-one (invertible)

Common nonlinear transformations

<table>
<thead>
<tr>
<th>Name</th>
<th>Functional Form</th>
<th>Inverse Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log</td>
<td>$g(R; \gamma) = \ln(R)$</td>
<td>$g^{-1}(r; \gamma) = \exp(r)$</td>
</tr>
<tr>
<td>Power</td>
<td>$g(R; \gamma) = R^{\gamma}$</td>
<td>$g^{-1}(r; \gamma) = r^{1/\gamma}$</td>
</tr>
<tr>
<td>Box-Cox</td>
<td>$g(R; \gamma) = (R^\gamma - 1) / \gamma$</td>
<td>$g^{-1}(r; \gamma) = (\gamma r + 1)^{1/\gamma}$</td>
</tr>
<tr>
<td>Box-Cox(\delta)</td>
<td>$g(R; \gamma) = [(R + \delta)^\gamma - 1] / \gamma$</td>
<td>$g^{-1}(r; \gamma) = (\gamma r + 1)^{1/\gamma} - \delta$</td>
</tr>
</tbody>
</table>
Building a statistical model for 24HRs

**Original Scale**

![Graph of Original Scale](image)

**Transformed Scale**

![Graph of Transformed Scale](image)

### Extended model for 24HRs

\[ T_i = E[R_{ij} | i] \]

\[ g(R_{ij}) = \mu + u_i + v_j \]

\[ u_i \sim N(0, \sigma_u^2), v_j \sim N(0, \sigma_v^2) \]

\[ \text{Corr}(u_i, v_j) = 0 \text{ for all } i \text{ and } j \]

---

### A warning about means of transformed data

- Averaging transformed data is **not the same** as transforming averages of raw data if the transformation is nonlinear.

\[ E[g(R)] \neq g(E[R]) \]

- Taylor series argument:

\[ E[g(R)] = g(E[R]) + \frac{1}{2} g''(E[R]) \text{Var}(R) + \text{extra terms} \]

- Extra terms involve “higher-order moments”

---

### Summary

- **Unbiased ≠ error-free**
- Within-person variation → overdispersion
- Model built using additional assumptions
  - Common variance components
  - Distributional assumptions (optional)
- Skewed distributions of intake may be handled with transformations

---

### Data requirements

- Two or more 24HRs on at least a subsample
- Replicate 24HRs should be far apart in time to maximize information
- Distribution of 24HRs should be "normalizable"
  - Unimodal, no spikes at extreme values
Two general approaches

- Model-Assisted (M-A) – rescales observed individual mean distribution
- Model-Based (M-B) – estimates distributions from theoretically-derived quantities

Rationale for the Model-Assisted approach

\[ R_i = \mu + u_i + \epsilon_i, \quad \text{Var}(u_i) = \sigma_u^2, \quad \text{Var}(\epsilon_i) = \sigma_\epsilon^2 \]

- For a sample of single 24HRs:
  \[ \text{E}[R_i] = \mu \quad \text{Var}(R_i) = \sigma_u^2 + \sigma_\epsilon^2 \]
- For a sample of \( J \)-day means:
  \[ \text{E}[\bar{R}_i] = \mu \quad \text{Var}(\bar{R}_i) = \sigma_u^2 + \frac{\sigma_\epsilon^2}{J} \]

Implementing the Model-Assisted approach

- Fit model to obtain parameter estimates
- Scale individual means to have desired variance

\[ r_i = \frac{(R_i - \hat{\mu}) \sqrt{\frac{\sigma_u^2 + \hat{\mu}}{\sigma_u^2 + \frac{\sigma_\epsilon^2}{J}}} \]

- Use empirical distribution of \( r_i \) as estimate of usual intake distribution

Features of the Model-Assisted approach
### Interpretation of scaled means

- The scaled means $r_i$ are not intended to be estimates of individual usual intake.
- The distribution of scaled means has the same mean and variance as the distribution of usual intakes in the population.
  - Distributions coincide for normal distributions.
  - Agreement only approximate otherwise.

### Features of the Model-Assisted approach

- Data-driven, uses few assumptions.
- Only requires separation of variance components.
- Precision of empirical percentiles limited.
  - There are only $N$ jumps in estimated distribution function.

### Estimating distributions from the model

#### Interpretation of scaled means

- The scaled means $r_i$ are not intended to be estimates of individual usual intake.
- The distribution of scaled means has the same mean and variance as the distribution of usual intakes in the population.
  - Distributions coincide for normal distributions.
  - Agreement only approximate otherwise.

### Rationale for the Model-Based approach

- Distribution of usual intake is specified by estimated model parameters:
  $$R_y = \mu + u_i + \epsilon_i,$$  
  $$u_i \sim N(0, \sigma^2_u), \quad \epsilon_i \sim N(0, \sigma^2_\epsilon)$$
- Probabilities/percentiles can be computed from tabulations of the standard normal distribution:
  $$\Pr(T \leq c) = \Phi\left(\frac{c - \hat{\mu}}{\hat{\sigma}_u}\right),$$  
  $$q_{\alpha}(T) = \hat{\mu} + \hat{\sigma}_u \Phi^{-1}(\alpha) = \hat{\mu} + \hat{\sigma}_u \Phi^{-1}(\alpha)$$

### Implementation using Monte Carlo simulation

- Randomly draw many (say $K$) values from the assumed normal distribution:
  $$u_k \sim N(0, \sigma^2_u)$$
- Create simulated usual intake (pseudo-value):
  $$r_k = \hat{\mu} + u_k$$
- Use empirical distribution of $r_k$ as estimate of usual intake distribution.
**Features of the Model-Based approach**

- Less robust, uses more assumptions than M-A
- Assumes distribution of \( u \) is known
- More precise percentile estimates
  - No limit to smoothness of estimated distribution function

**Accounting for nonlinear transformations**

- Goal is to estimate a quantile of usual intake that corresponds to one in the normal distribution that exhibits only between-person variance

**Estimating quantiles when transformations are used**

- With no transformation used:
  
  \[
  q_{p}(r) = E\left[\mu + u + \varepsilon \mid u = q_{p}(\mu)\right] \\
  = \mu + E[u \mid u = q_{p}(\mu)] + E[\varepsilon \mid u = q_{p}(\mu)] \\
  = \mu + q_{p}(\mu) 
  \]

- Estimated quantile is a linear function
Estimating quantiles when transformations are used

- With nonlinear transformation \( g \) used:
  \[
  q_{p(T)} = E[g^{-1}(\mu + u + \varepsilon)] \quad u = q_{p(T)} \]
  \[
  = E[g^{-1}(\mu + q_{p(T)} + \varepsilon)]
  \]

- Estimated quantile is an integral

- Can be calculated/approximated several ways

Integration provides the “backtransformation”

- Taylor series approximation (Dodd, 2006):
  \[
  q_{p(T)} \approx g^{-1}(\mu + q_{p(T)} + \frac{1}{2}(g^{-1})'(\mu + q_{p(T)})\varepsilon^2)
  \]

- Exact calculation for normal \( \varepsilon \) (Hoffmann, 2002)

- Numerical integration for known \( \varepsilon \) distribution
  - Quadrature formulas, e.g., Gauss-Hermite
  - Monte Carlo integration

Estimation approaches when transformations used

- Both Model-Assisted and Model-Based approaches can be extended

- If transformation \( g \) achieves the desired distribution of \( \varepsilon \) terms, Taylor series approximation may be poor
  - Alternatives use all moments, not just two

Evolution of estimation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Transformation</th>
<th>Distributions via</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC (1986)</td>
<td>None*</td>
<td>M-A</td>
</tr>
<tr>
<td>Slob (1993)</td>
<td>Log</td>
<td>M-B</td>
</tr>
<tr>
<td>BP (1996)</td>
<td>Power</td>
<td>M-A</td>
</tr>
<tr>
<td>ISU (1996)</td>
<td>Two-stage</td>
<td>M-B/M-A</td>
</tr>
<tr>
<td>NCI (2006)</td>
<td>Box-Cox</td>
<td>M-B/M-A</td>
</tr>
<tr>
<td>MSM (2011)</td>
<td>Box-Cox</td>
<td>M-A</td>
</tr>
<tr>
<td>SPADE (2012?)</td>
<td>Box-Cox</td>
<td>M-B</td>
</tr>
</tbody>
</table>

* NRC method incorporates transformations under alternative assumptions

Software availability for estimation methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Software?</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC (1986)</td>
<td>Yes</td>
<td>SAS/C/Windows</td>
</tr>
<tr>
<td>Slob (1993)</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>BP (1996)</td>
<td>Yes</td>
<td>SAS/C/Windows</td>
</tr>
<tr>
<td>ISU (1996)</td>
<td>Yes</td>
<td>SAS/C/Windows</td>
</tr>
<tr>
<td>NCI (2006)</td>
<td>Yes</td>
<td>SAS</td>
</tr>
<tr>
<td>MSM (2011)</td>
<td>Yes</td>
<td>R (via Website)</td>
</tr>
<tr>
<td>SPADE (2012?)</td>
<td>Yes (beta)</td>
<td>R</td>
</tr>
</tbody>
</table>

Summary

- Within-individual variation is adjusted out, leaving only between-individual variation

- Two approaches to estimate distributions
  - Model-assisted vs. Model-based

- Use of normalizing transformations requires special care in estimating distributions
  - Backtransformations of varying complexity

- Wide range of software implementations
The role of covariates

The need for subpopulation estimates

- Nutritional status often depends upon personal characteristics

One answer is to stratify sampled data

- 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
The role of covariates

Example: Eating at America’s Table Study (EATS)

Estimating usual intake distributions for dietary components consumed daily by nearly all persons

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

A mixed model formulation

\[ R_i = \mu + u_i + \epsilon_i \]

- Population mean is a fixed effect
  - Only one model parameter to estimate

- Deviations are random effects
  - Reflect variation from individual persons/days
  - Focus on higher-order moments, e.g., variance

- Mixed models include fixed and random effects

A mixed model formulation including covariates

\[ R_i = \mu(X) + u_i + \epsilon_i \]

- Fixed effect part of the model expressed as a function of measured covariates X
  - Multiple parameters to estimate
  - Allows “structured” variability in group means

- Random effects reflect variation from all other unmeasured characteristics
  - "Unstructured" variability

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)
Allows different means for subpopulations, while pooling information about variance components
- Point estimates for overall population may be unaffected by covariates,
- But should be more precise if model holds

Potential benefits of incorporating covariates

- Can investigate multiple determinants of diet
  - Test significance of main effects/interactions
  - Joint modeling leads to lower analysis burden

Overall bias due to nuisance effects can be corrected

In epidemiologic applications, less unstructured variation is better

Estimating distributions with covariates in the model

Model-Assisted: use observed covariate pattern $X_i$ for $i$-th individual:
$$ r_i = (\bar{R}_u - \hat{\mu}(X_i)) \left[ \frac{\sigma^2_{\mu}}{\sigma^2_{\mu} + \frac{\hat{\sigma}^2}{J}} \right] $$

Model-Based: use a specified covariate pattern $X_0$ for $k$-th pseudo-value:
$$ r_k = \hat{\mu}(X_0) + U_k $$

Model-Assisted and Model-Based similar unless
- Important covariate(s) are omitted, and/or
- Exact normality does not hold

Discrepancy between Model-Assisted and Model-Based distributions useful as a diagnostic

Discrepancy between Model-Assisted and Model-Based distributions useful as a diagnostic

Model-Based approach also offers a choice to perform direct standardization
- Draw $X_0$ from a standard population
The role of covariates

Direct standardization for time-dependent covariates

- Overall usual intake is weighted average of time-dependent usual intake

- Weights come from the standard population, e.g., for weekend/weekday effects.

### Standard Population for Weekdays/Weekend Days

<table>
<thead>
<tr>
<th>Weekend</th>
<th>Days of Week</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>No</td>
<td>MTWT</td>
<td>4/7</td>
</tr>
<tr>
<td>Yes</td>
<td>FSS</td>
<td>3/7</td>
</tr>
</tbody>
</table>

The role of covariates

Explicit adjustments for nuisance effects

- Can be done before fitting the mixed model, or

- In a two-stage process:
  - Include nuisance effects in the mixed model
  - Estimate distributions using group means calculated with nuisance covariates set to fixed reference values, e.g., the first interview, or the in-person interview

The role of covariates

Types of covariates allowed in available methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Covariates Allowed</th>
</tr>
</thead>
<tbody>
<tr>
<td>NRC (1986)</td>
<td>None</td>
</tr>
<tr>
<td>Slob (1993)</td>
<td>None</td>
</tr>
<tr>
<td>BP (1996)</td>
<td>Nuisance</td>
</tr>
<tr>
<td>ISU (1996)</td>
<td>Nuisance</td>
</tr>
<tr>
<td>NCI (2006)</td>
<td>Individual, Time-dependent, Nuisance</td>
</tr>
<tr>
<td>MSM (2011)</td>
<td>Individual</td>
</tr>
<tr>
<td>SPADE (2012?)</td>
<td>Individual, Time-dependent*, Nuisance*</td>
</tr>
</tbody>
</table>

* fractional polynomial option for age

The role of covariates

Summary

- Covariates provide an alternative to stratification

- Mixed model allows a combination of structured and unstructured variation

- Both approaches to distribution estimation (M-A and M-B) can be extended to handle covariates of three types: individual, time-dependent, and nuisance

- Not all available methods incorporate covariates; if they do, implementations vary

QUESTIONS & ANSWERS

Moderator: Sharon Kirkpatrick

Please submit questions using the Chat function

Next Session
Tuesday, October 4, 2011
10:00-11:30 EDT

Estimating usual intake distributions for foods and nutrients consumed episodically

Janet Tooze
Wake Forest University