Ethiopia Poverty Measurement Training

Day 1 & 2 data cleaning

Data Cleaning Goals

1. Identify and address systematic mistakes in the data (resulting from mistakes in
2. data entry program, data processing, widespread misunderstanding of the questionnaire etc)
3. Address missing values so they don’t unintentionally result in loss of precision or downward bias in consumption
4. Identify and address outliers without systematically trimming the tails of the distribution, especially gross outliers that would influence the poverty status of a hh
5. Ensure random errors do not have undue influence on results (through construction of key models, prices etc)

Systematic Mistakes – The Problem

- A variety of processes (mistakes in questionnaire design, programming of data entry application, data processing, misunderstanding of the questionnaire, instrument failure) can result in systematic mistakes which mean your data is not what it should be, not what you think it is based on the questionnaire.
- Do not assume your data is what it should be.
- This has the potential to be hugely influential on your results.
- Solution: inspect your data carefully and look carefully at the distribution of all aggregates and sub-aggregates.

Missing Values – The Problem

- Constructing the consumption aggregate requires aggregating information from hundreds or thousands of variables across thousands of households.
- Depending on how this is aggregated, missing values are either implicitly treated as zero (if using commands like `collapse` or `egen rowtotal()`) or result in a missing value for the aggregate (if using the + operator).
  - The first results in the total consumption being underestimated, potentially by enough to impact the poverty rate.
  - The second will result in the consumption aggregate not being constructed for a significant number of households.
- Even worse, say the construction of sub-aggregates (like food consumption) results in a missing if any component variable is missing, but the construction of the total aggregate treats missing food consumption as 0.
- Solution: be very thorough and deliberate in your treatment of missing values.

Outliers – The Problem

- A single gross outlier (such as rice consumption being reported as 2000 kg when it is really 2000 g) can result in a household having a level of consumption above the poverty line due solely to that one mistake.
- Errors of several orders of magnitude are quite possible due to sticky-0 key, mistakes over units, mistakes about recall periods, mistakes about currency conventions (amounts to be entered sometimes in Leones, sometimes in thousands of Leones).
- Solution: identify and treat gross outliers.

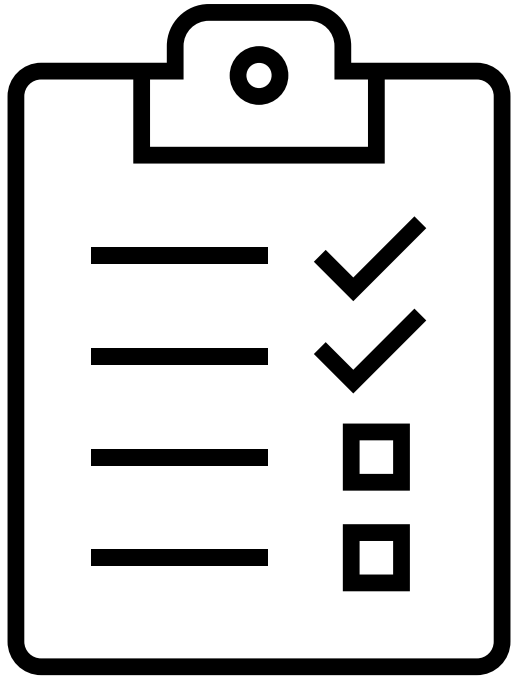
Outliers – Not (Really) a Problem

- Values of variables are randomly 50% higher or even twice as high as they should be.
- Values at the top end of the distribution are a little suspicious although plausible for wealthy households.
- There are implausible values at the bottom end of the distribution (households spending amounts less than the smallest denomination bill/coin, less than smallest quantity of the good its possible to purchase).
- Although ideally we would like to identify and correct for all of these, it is very difficult (to impossible) and very time-consuming to do.
- As long as these types of errors are fairly rare and fairly random, they will have basically no impact on the poverty rate.
- They might impact measures of inequality like the Gini (but general household questionnaires aren't designed to capture the consumption in the tails of the distribution anyway)
- Problem: over trimming/winsorizing your data can artificially compress the distribution
- Solution: do a reasonable job on moderate outliers without wasting too much time or over-trimming/winsorizing.

Random Errors – The Problem

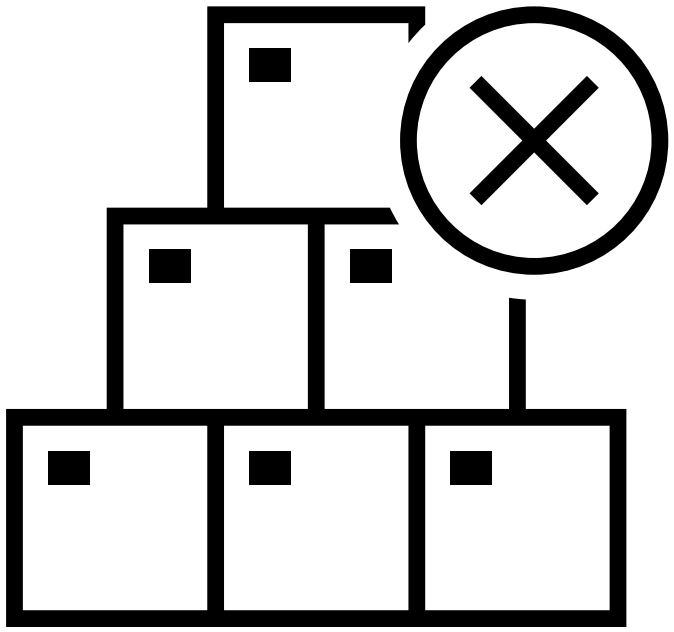
- A single gross outlier or a cluster of outliers could result in an erroneous construction of a key value (such as the regional price for a key staple food) or an erroneous model (such as the imputation model for the value of owner-occupied housing).

- Solution: use robust constructions (median instead of mean etc) and be conservative about the observations used as input. Check the results of any constructions before applying them to the data.



Three Step Approach

1. Domain, obvious and systematic editing of each dataset that is to be used.
2. Comprehensive treatment of missing values and outliers
    a. In all variables used
    b. In all variables constructed
3. Check of distribution of final aggregates



Step 1: Domain, obvious and systematic editing

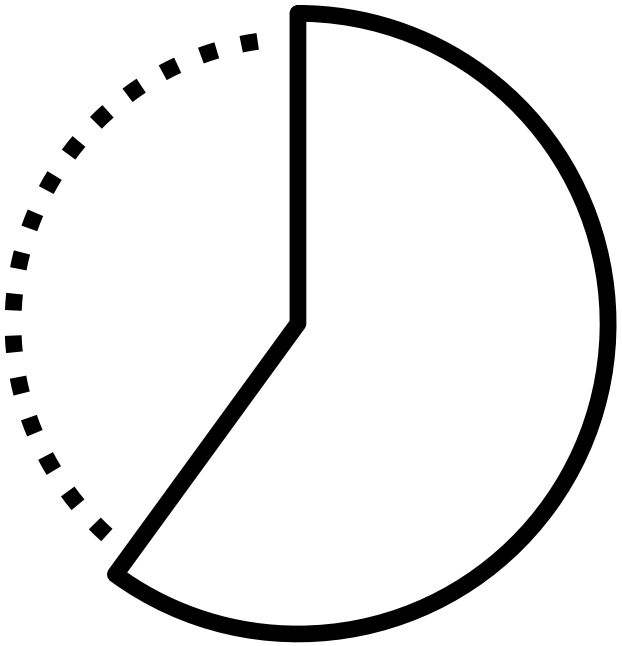
- Trust no one
- Every time you use a dataset, inspect it carefully

Domain / Obvious / Systematic Errors

- Check all inputs into your constructions
- Check for invalid codes
- Check of invalid combinations of items and units
- Values that are several orders of magnitude wrong / not just improbable but impossible.
  - It's improbably that anyone in our surveys would pay $10,000/month on rent. It's impossible that they pay $10,000,000/month
- Outliers and missing values scattered at random are pretty harmless.
  - High levels of outliers or missing, particularly if they are clustered in a specific item / area / interviewer / period can be very problematic
- In future steps dealing with missing and outliers, always be on the lookout for systematic errors

Domain / Obvious / Systematic Errors

- Only worry about extreme outliers or patterns of outliers here
- Don’t change the existing data.
  - Identify issues that need to be investigated / addressed further
  - Flag observations to exclude from subsequent constructions
- We will address low levels random-ish missing values and moderate outliers in a more systematic manner later



Step 2: Comprehensive Treatment of Missing Values and Outliers

- Identify
- Address

Flagging missing and invalid observations

- Missing and invalid observations can include:
  1. Actual missing values (when the question should have been asked)
  2. Invalid values (such as negative values or truly impossibly large values)
  3. Invalid zeros (the household reported spending on the item, but the expenditure amount is given as 0)
  4. Amounts that are considered too low to be a plausible expenditure transaction for the item (e.g., it's not possible to buy $0.05 worth of fuel)
  5. Amounts that are not a valid currency amount for other reasons (not multiples of currency denominations in circulation)
  6. Codes for "don't know" or "refused" (either explicitly provided on the questionnaire, or values the interviewers likely used to indicate this such as 99 or 88)
- Flag all of these observations so that we can:
  - Impute a reasonable value of consumption based on the median instead
  - Exclude them from the identification of outliers

Identifying Missing Data

- For every component of the consumption aggregate, should there be data for this item / person / household?
- Sometimes there is a filter question
  - Did this child go to school? If yes, how much was paid in school expenses?
  - Does the household own a TV? If yes, what is the current value?
  - Did the household purchase soap in the last month? It yes, how much was spent to buy soap?
- Otherwise, the question is always asked and a 0 is recorded if nothing was spent.
  - How much did the household pay for water in the last month?

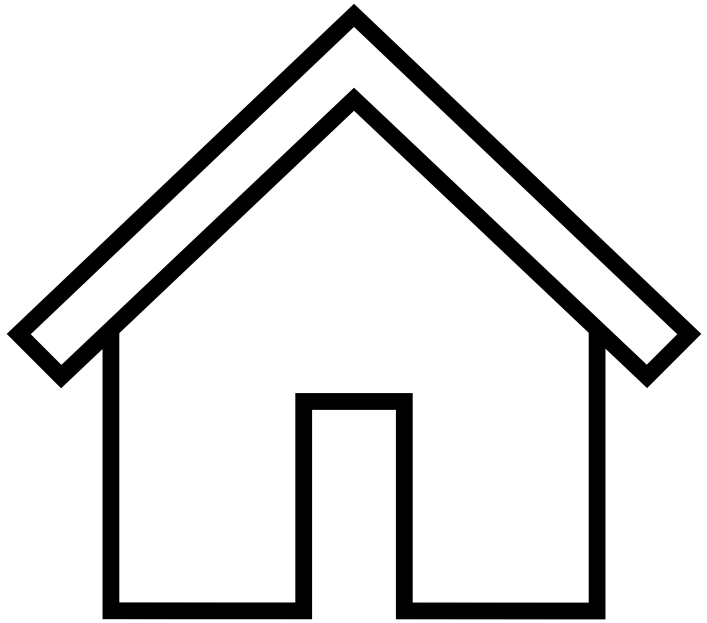
Identifying Missing Data

- If the question should be answered, is it? If not, that is a missing that we will have to address.

- What if the filter question is missing? Can try to impute that: do more than half of similar households report buying soap? (Define “similar” usually by geographic characteristics.)

Replacing Missing Values with Imputation

- Easiest / most common method:
  - Replace with [weighted] median value of variable for the relevant group
  - For total household expenditures, use value per capita (or other adjustment you are using for household size)
  - Geographic area and urban/rural (as these reflect price levels)
  - Anything else relevant. For example, for school fees, level of school (primary vs. secondary) and owner (public vs. private)
  - Consider households interviewed around the same time if inflation / seasonal price variations are a concern
- More sophisticated methods are available, see M&V 7.2/7.3 for discussion and references



Excessive Missing Values

- If too many items are missing for a household, may just want to drop the household from the analysis of poverty and adjust the weights considering this household as a non-response
- Rule of thumb used for Sierra Leone: drop households missing more than ~ 10% of data from any section

What is an Outlier?

- Not all outliers are errors. Not all errors are outliers.
- What we want to find and adjust for are **errors**, looking at outliers is one tool
- *"An outlying observation, or 'outlier,' is one that appears to deviate markedly from other members of the sample in which it occurs."* (Grubbs 1969)
- An outlier is not the same as the tail of the distribution. Every distribution has a tail.
- Avoid any rule that would always flag some observations as "outliers" even if the data was perfectly correct
  - Using too low a Z-score cut-off (consider expected number of observations with a given Z-score for your sample size)
  - Using any percentile cut-off
- If a variable is known to have a normal distribution, for any given sample size you can estimate how many observations you expect to have (say) more than 3 standard deviations from the mean.

Detection of Outliers

- Done on variable for expenditure / use value / value of consumption by item
- We generally assume per capita consumption / expenditure of each item is log-normally distributed
- Often done separately for subgroups of households by location or month of survey.
  - Do not make these groups too small.
  - Do not subdivide unless there is good evidence that expenditure patterns are different.
  - Often, urban/rural is sufficient.

Formulas to Identify Outliers

- Identify location (central tendency) of the distribution, a scale to measure distance from that center and set a threshold (maximum allowed distance)
- Location: mean (sometimes after trimming top and bottom of distribution) or median
- Scale: standard deviation, mean absolute deviation (MADS) from the median (less influenced by extreme values), interquartile range etc
- Threshold: can be set at values corresponding to a Z-score of 2.5, 3, 3.5 etc. (Recommend at least 3.)

Examples of Formulas to Identify Outliers

- **Classic Z-score**: uses mean as location, standard deviation as scale. Distance from center is expressed as Z-score, so threshold is expressed as a Z-score of ±3 etc.
  - Concerns: estimation of mean and standard deviation are easily biased by the presence of extreme values

- **Robust Z-score** (as described by M&V): uses median as location and MADS as scale. Using $Z = (X - \text{med}(X))/(\text{mads}(x)/0.675)$ gives something equivalent to Z-score in that, if X is normal, same percent of distribution has Z-score < 3.

Examples of Formulas to Identify Outliers

- Tukey: Usually expressed as [Q1 – n * IQR, Q3 + n * IQR] where n = 1.5 to 2.5. Uses median as location (assuming symmetric distribution), IQR as scale. Z = (X – med(X))/(iqr(x)/1.35) is equivalent.

- Alternative: Drop top and bottom 5% of distribution and take mean as location. IQR as scale.

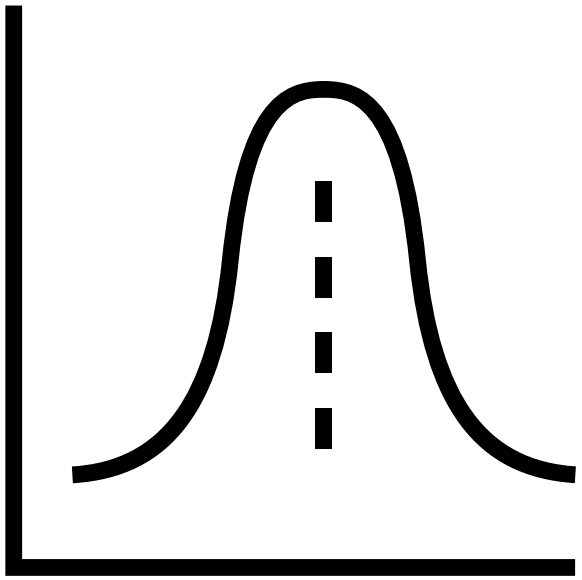
- All of these are fine and basically equivalent.

Lower Outliers

- Includes inappropriate values of 0 (household reports consuming but 0 for amount spent)
- Often best analyzed based on total household consumption (as opposed to per capita/per adult equivalent for upper outliers)
- Consider smallest possible quantum that can be purchased
  - How much for one book of matches? One cup of rice?
  - What is the smallest unit of currency in circulation?
- Very often just use smallest currency denomination

Treatment of Outliers Believed to be Errors

- Options
  - Consider them as missing and replace with imputed value – see discussion above on missing at random
  - Winsorize – replace with largest (smallest) non-outlier value.
    - Assumption is that household did purchase an unusually large amount, just not as much as recorded.
    - Or, more pragmatically, splitting the difference between believing the recorded value and not
- If analysis is done on per capita basis, make sure to multiply back by household size



Overall Distribution of Consumption

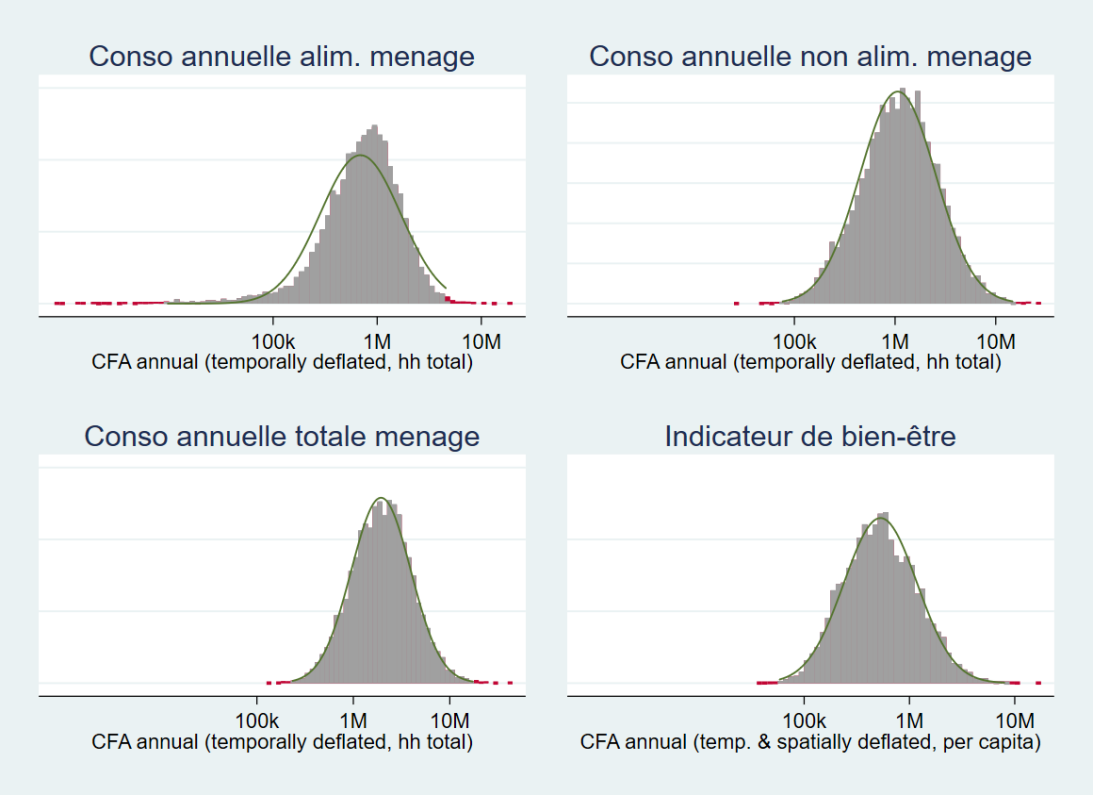
- Check the overall distribution of consumption
- Should be roughly log-normal

Checking Overall Distribution

- Check each component of the consumption aggregate as you construct it, both in total and per capita terms
- Check the distribution of total food consumption and total non-food consumption
- Any marked deviation from a log-normal distribution is cause for concern and should be investigated
  - Decompose by type of consumption: food into starches, meat, fruits/vegetables etc.
  - Decompose by type of household: farmer vs. not, renters vs. owners
- Check distribution of final welfare measure

Use of `outdetect`

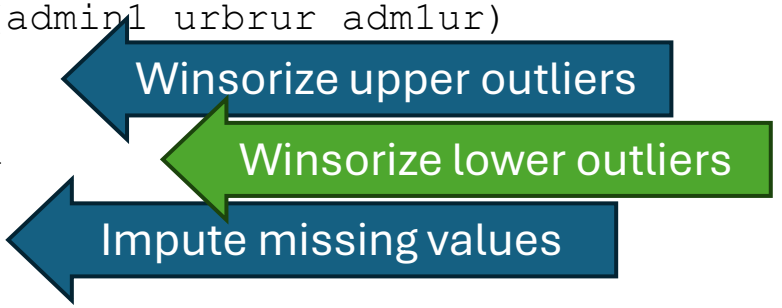
- Stata command written by M&V & coauthors
- Not robust enough / too slow to use in loop on every single item
- Very nice for looking at distribution of final aggregates



`flagout` example for outliers and imputation

- Consider expenditure data on general nonfood items
  - *b0* is item code
  - *b2* is amount spent in last month
  - missing and invalid values in *b2* have already been identified and flagged via the variable *miss_inv*
  - assume per capita expenditure is log-normally distributed

```
gen consexp = b2 if !miss_inv
gen lpccons = log(consexp/hhsize)
gen adm1ur = admin1 * 10 + urbrur
flagout lpccons [pw = hhweight], item(b0) z(3.5) over(admin1 urbrur adm1ur)
replace consexp = hhsize * exp(_max) if _flag == 1
// replace consexp = hhsize * exp(_min) if _flag == -1
replace consexp = hhsize * exp(_med) if miss_inv
```



`flagout` syntax

```
syntax varname [pweight] [if], item(varlist max=1) [over(varlist)
   minn(integer 30) z(real 3.5)]
```

- *varname* : variable whose distribution is to be analyzed, assumed to be normally distributed
  - transform variable (make per capita, take logs) before passing to program
- Sampling weights are allowed (and recommended to be used [currently required])
- Distribution of values of *varlist* for each value of *item* are assumed independent.
  - required, so construct constant variable if needed
- Over allows ordered set of variables to define subgroups
  - the **last** variable in the list with min number of obs will be used to define the location and scale
- minn sets the minimum number of observations required; 30 is the recommended default
- z specifies the Z-score beyond which observations are considered to be outliers and flagged

`flagout` details

- Uses median as center of distribution
- Uses interquartile range (p75 – p25) as scale if this is non-zero, otherwise p90 – 10
  - these are rescaled to make the Z-score thus constructed equivalent to the standard Z-score (constructed with the standard deviation) on a normal distribution
  - if p10 = p90 (that is, at least 80% of the observations take the same value), any observation not equal to this value will be flagged as an outlier
- Constructs four variables:
  - _flag indicates if an observation is identified as a lower outlier (-1), upper outlier (1) or nonoutlier (0)
  - _min is the lower limit of non-outliers
  - _max is the upper limit of non-outliers
  - _med is the median
- Each is defined over the relevant subgroup (last *over* variable with at least *minn* observations)

Data clean – step1

Variable names

- Keep the original name or change to a new one

Labels

- Variable label: Make sure all the variables have a valid description.
- Make sure all the value labels are correctly defined
  - Yes: 1 No: 0
  - Urban: 1 Rural: 2

Duplicates check

- Real duplicates
- Enter error
- Keep or drop

Missing check

Data clean – step2

Validate the skipping pattern

- Some questions will be omitted for some households or individuals if a specific condition is met.
- eg. if “others” are selected, the following question will be asked to specify

Validate the data

- Checking missing
- categorical variables: if the value within the range
- String variables: if all the input meaningful. Especially for “others”

Data clean – step3

Issue fix

- Some common problems:
  - Wrong unit entered. 2,000g instead of 2,000kg
  - Missing decimals: 200000 instead of 2000.00
  - Currency issues

Hard code

- Use it when necessary
- We should try not to change the value directly to avoid manipulation
- Some outliers could be fixed by reviewing the data by the statistical team.

Other extra topics

Household Weight

When to use household weight

Use it when the statistic you want describes **households as units**.

Examples:

- % of households with electricity
- Average household expenditure
- Distribution of household sizes
- Ownership of durable goods at the household level (e.g., % of households owning a car)
- Here, each household counts once, regardless of how many people live in it.

Individual Weight

When to use population (individual) weight

Use it when the statistic you want describes **individuals (people)**, or when the measure is **per capita**.

Examples:

- Poverty headcount rate (% of people living below $2.15/day)
- Literacy rate of adults
- Labor force participation rate
- Average consumption **per person**
- Demographic breakdown (age, sex, etc.)

Here, households with more members contribute more because they represent more people in the population.

weight comparison

The intuition:

Household weights → “Each household represents many similar households.”

Population weights → “Each person represents many similar people.”

Since households vary in size, the two weights give different results if you mix them up.

For example:

- If you estimate poverty with **household weights**, you’d get the **share of households** that are poor.
- If you use population weights, you get the share of people who are poor (this is the standard poverty rate)

Rule of thumb

Unit = household → use weight_hh

Unit = individual or per capita → use weight_pop

Extra data needed

- Conversion factor:
  - Change form the not standard unit(heap, tin, etc.) to standard unit (kg/l)
- Market unit price for food
  - Ex: Average price for rice in kg
- Energy intake conversion factor
  - Food energy for per 100g items