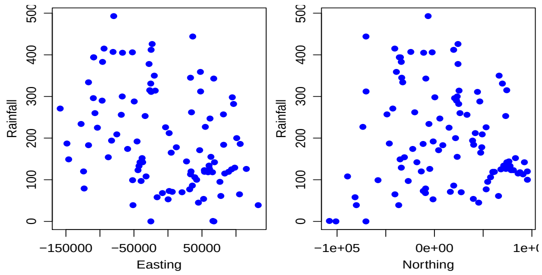
Visualizing spatial data

Goal: draw a picture to illustrate interesting patterns in data

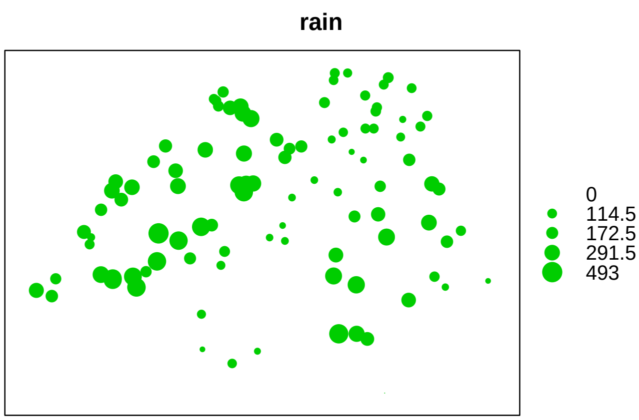
- Problem: Spatial data is 3D: X,Y for location and Z for value
- Could plot Z vs X and Z vs Y: incomplete

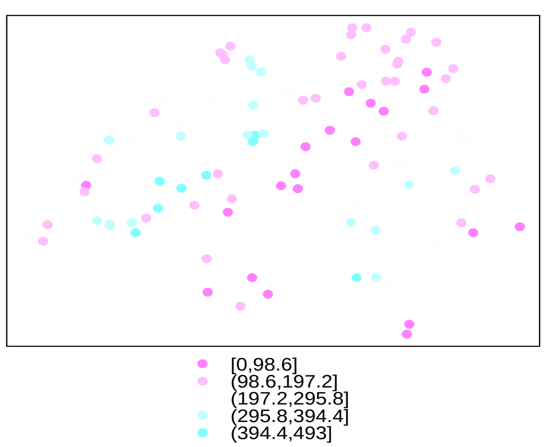


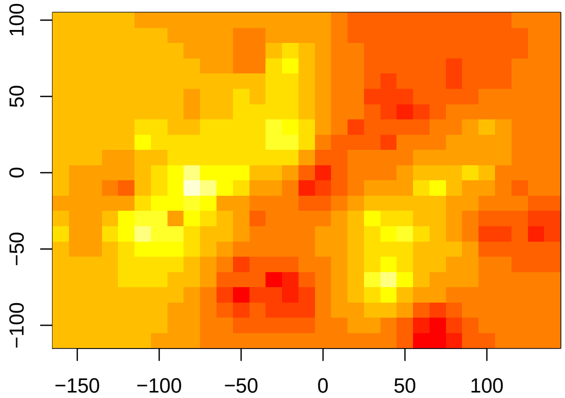
Visualizing spatial data

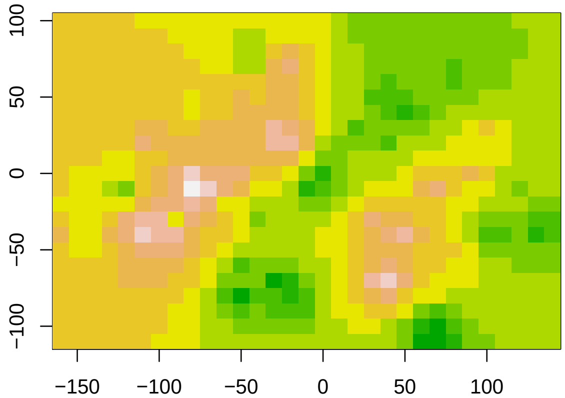
Many different solutions. I'll illustrate various

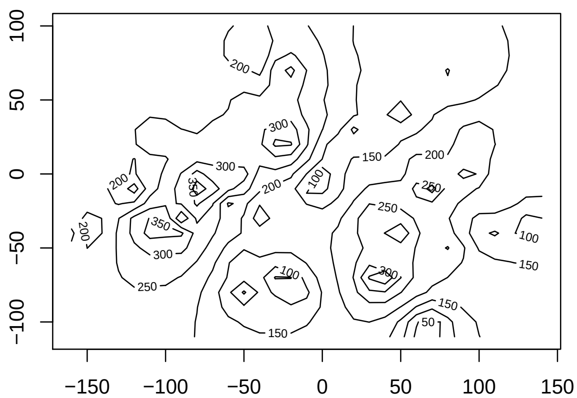
- bubble plot: radius of symbol proportional to $\sqrt{Z}$
  Avoids a graphical illusion: we see area, not radius/diameter
  so radius $\propto \sqrt{Z}$ means area $\propto Z$.
- colored dot plot: color indicates Z
- image plot: color indicates Z
- contour plot: lines indicate Z
- Avoid perspective plots. They usually don't work well.
- more focused plots for specific situations
  - Conditioning plots: compactly show subsets of data all at once
  - Z vs X for bands of Y
  - Spatial plot for each time

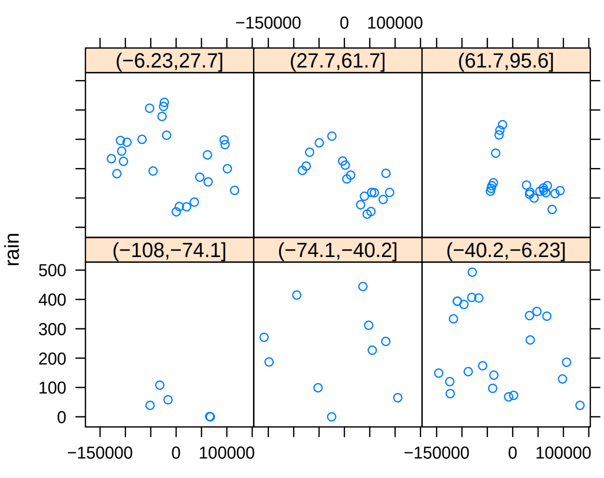


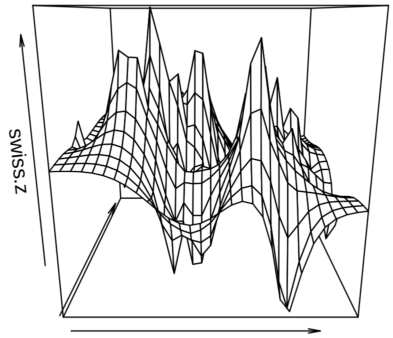



100
50
0
−50
−100
−150
−100
−50
0
50
100


100
50
0
−50
−100
−150
−100
−50
0
50
100

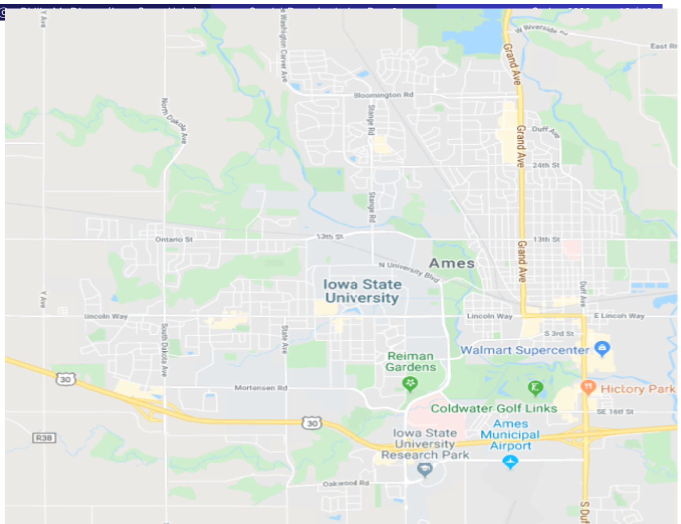

200
300
300
150
200
200
250
250
100
200
350
300
350
250
300
250
100
100
150
150
150
50
100
200
−150
−100
−50
0
50
100
150

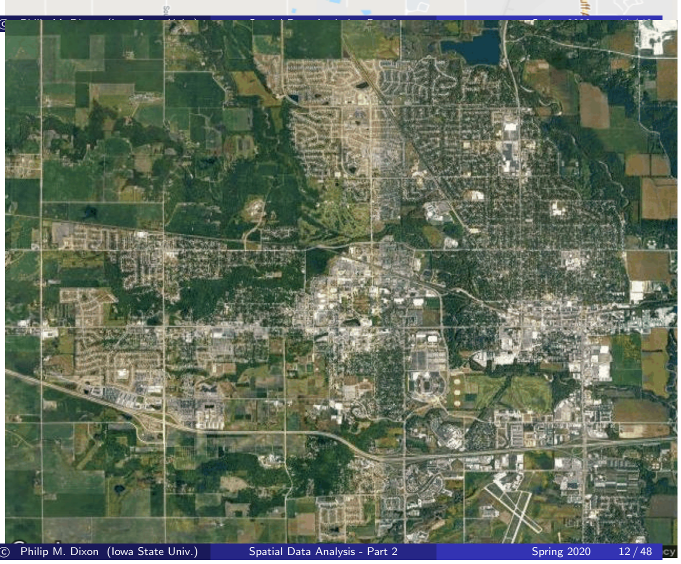

−150000 0 100000
(−6.23,27.7]
(27.7,61.7]
(61.7,95.6]
(−108,−74.1]
(−74.1,−40.2]
(−40.2,−6.23]
rain
x
0
100
200
300
400
500
−150000 0 100000

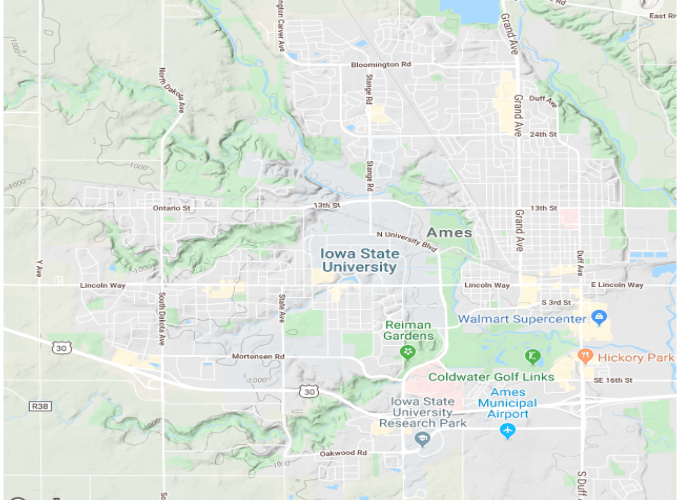


Visualizing data on maps

- You can plot data on maps
  - Can show just the locations or values at locations
  - Often more informative than on a blank background
- Need to get the map: 3 major sources
  - OpenStreetMaps: appears to not be available right now
  - Bing: requires registration and an API key
  - Google: street maps are open, other images require API key



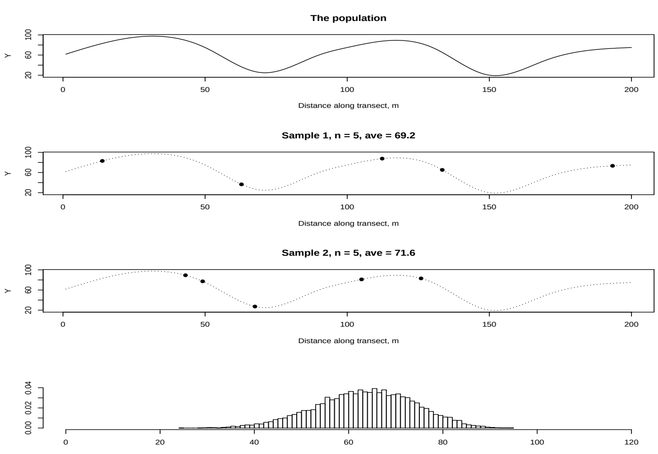




Spatial sampling

Consider a population of 2000 objects along a line (next slide)

- Want to learn about this population, but can only afford n =5 samples
- Draw a sample, estimate sample quantities, infer to the population
- Simple random sample
  - usually without replacement
  - every unit has same probability of occuring in the sample
    - inclusion probability
  - every pair of units has the same probability of occuring together in the sample
    - joint inclusion probability



Simple random sample

Simpler population: 2000 students

- randomly select and measure 5 of the 2000 units in the population
- calculate sample average: $\bar{Y} = \frac{\sum Y_i}{n}$
- and sample variance: $s^2 = \frac{1}{n-1} \sum \left(Y_i - \bar{Y} \right)^2$
- and se of $\bar{Y} = \sqrt{s^2/n}$

Questions:

1. Why is $\bar{Y}$ a good way to estimate μ?
2. Why is s^2 a good way to estimate σ^2?
3. Is $\sqrt{s^2/n}$ a good estimate of the variability of $\bar{Y}$?

Why do the usual things?

Why "usual" quantities are good quantities:

1. because of theoretical properties of the estimators.
2. because of simulation studies do not uncover problems.

Notation / vocabulary:

- E X is the expected value (theoretical average) of X, a random variable.
- Estimator: The function that computes an estimate
- Estimate: A value for a specific data set

Properties of estimators

Results from 587/588 stated / proved using a model for the data

$$Y_i = \mu + \varepsilon_i, \;\; \varepsilon_i \sim \text{iid } N(0, \sigma^2)$$

- Observations (or errors) are independent
- With constant variance
- And mean error = 0 for all observations

Why the mean is good:

- Unbiased: E $\bar{Y} = \mu$
- Minimum variance among unbiased estimators for this model:

(optional) Proof / elaboration

Why mean is unbiased

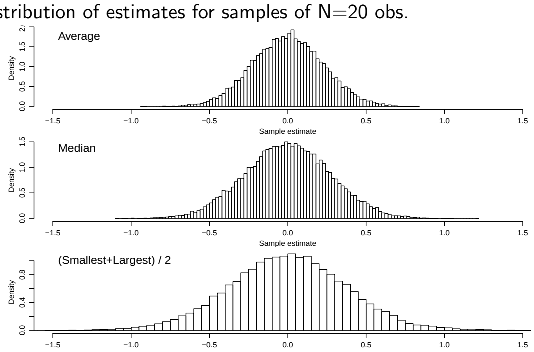
- Model says E $\varepsilon = 0$, so E $Y = \mu$
- E $\bar{Y} =$ E $[\Sigma Y_i / n] = \Sigma$E $[Y_i] / n = \Sigma \mu / n = n\mu / n = \mu$

Minimum variance among all unbiased estimators

- $\bar{Y}$ is a random variable. Estimates μ. Has a variance.
  Note Var $\bar{Y}$= se^2
- Consider another unbiased estimator of μ. Call it θ. E $\theta = \mu$.
- Can prove: Var $\bar{Y} \leq$ Var θ
- True for **any** θ that is unbiased
- $\bar{Y}$ better (or never worse) than any other unbiased estimator of μ.
- When Var θ is the criterion for better

Properties of estimators

- Example: Model above (Normal errors). Three ways to estimate μ: average, median, and mid-range: ave. of smallest and largest value.
- Distribution of estimates for samples of N=20 obs.



Properties of estimators

- Example: Model above (Normal errors) with $\mu = 0$.
- Numeric summaries of the three sampling distributions

	Sampling	
Statistic	average	sd
average	0.00	0.224
median	0.00	0.272
mid-range	0.00	0.378

- All three estimators are unbiased:
  - Population mean: 0.0000. All estimators are 0.00, on average
- Sample average is the least variable.
- Sample variance is an unbiased estimate of σ^2
- And s^2/n ($=$ se^2) is an unbiased estimate of the variance of $\bar{Y}$

Properties of estimators

- Above result seems obvious:
  - sample average uses all the observations, so isn't it obviously the best?
- Not at all a duh, obvious.
- New model for data: uniform distribution: $Y_i \sim U(a, b))$
- a and b not known, $\mu = (a + b)/2$.
  - Assume population is $U(0, 2)$, $\mu = 1$
- Best estimator of μ is now the mid-range.

Statistic	mean	se
average	1.00	0.13
median	1.00	0.21
mid-range	1.00	0.066

- Crucial point: "good" or "not-so-good" depend on the model

Back to Simple Random Sample of a transect

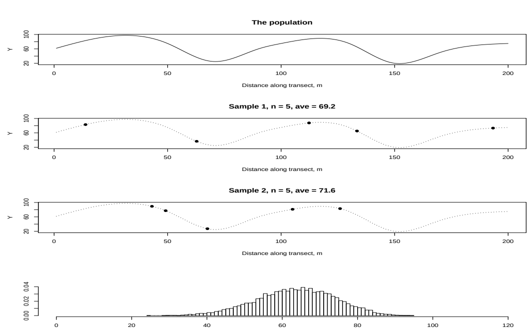
Measure 5 locations along our transect. Simple random sample.

- Randomly choose locations to measure.
- All locations are equally likely to be chosen
- SRS: all sets of 5 locations equally likely to be chosen
- Analyze in usual way: $\bar{Y}$, s, se of $\bar{Y} = s/\sqrt{n}$

The population has a clear spatial trend.
Units are similar to their neighbors
Questions:

1. What can we say about $\bar{Y}$? Is it still good?
2. Are the sample average and sd still valid estimators of the population quantities?
3. Is that se calculation still appropriate?

Back to a Simple Random Sample



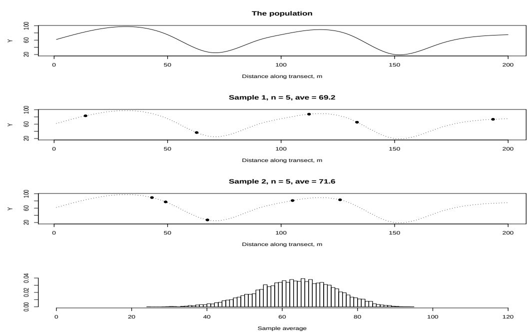
Answers

- What can we say about $\bar{Y}$? Is it still good?
- Are the sample average and sd still valid estimators of the population quantities?
- Is that se calculation still appropriate?

- A: Yes, to all questions.
  Spatial correlation in the population does not make usual estimators "bad"
  But, often can use spatial correlation to get a better estimator

Question

- Q: When you sample from a population, what is the random variable?



Answer

- Q: What is the random variable?

- A: It is not the value attached to a population unit, Y_i.
  - The Y_i are assumed to be fixed values, one for each unit.
  - The value for unit 125 doesn't change because it was or wasn't sampled.
- The only random variable in the classic approach to sampling is whether or not the i'th unit is included in the sample.
- Example of design based inference
  - Statistical conclusions justified by how the data were collected
  - not by an imaginary model (model based inference)
- Huge practical consequences.

(optional) Properties of SRS for spatial data: average

- Define $S_i = I(\text{unit } i \text{ is in the sample})$
- E $S_i = \frac{\Sigma S_i}{N} = \frac{n}{N} =$ P[unit i in the sample]
- $\bar{Y} = \frac{\Sigma_{all\ obs} S_i\ Y_i}{N}$
-

$$\begin{aligned} E\ \bar{Y} &= \frac{\Sigma_{all\ obs} S_i\ Y_i}{n} = \frac{1}{n} E \Sigma_{all\ obs} S_i\ Y_i = \frac{1}{n} \Sigma_{all\ obs} Y_i\ ES_i \\ &= \frac{1}{n} \Sigma_{all\ obs} Y_i\ \frac{n}{N} = \frac{\Sigma_{all\ obs} Y_i}{N} = \mu \end{aligned}$$

(optional) Properties of SRS for spatial data: variance

- Usual expression for s^2 is clumsy to work with
- Another formula for the sample variance is $s^2 = \frac{\Sigma_{j>i}(Y_i - Y_j)^2}{n(n-1)}$
  Try it sometime!
- Define $S_{ij} = I$(sample includes units i and j)
- $\text{E } S_{ij} = \frac{\Sigma_{j>i} S_{ij}}{N(N-1)/2} = \frac{n(n-1)/2}{N(N-1)/2} = \frac{n(n-1)}{N(N-1)} =$
  P[units i and j in the sample]
-

$$\begin{aligned} \text{E } s^2 &= \text{E } \frac{\Sigma_{j>i}^N S_{ij}(Y_i - Y_j)^2}{n(n-1)} = \frac{\Sigma_{j>i}^N (Y_i - Y_j)^2 \text{E } S_{ij}}{n(n-1)} \\ &= \frac{\Sigma_{j>i}^N (Y_i - Y_j)^2}{n(n-1)} \frac{n(n-1)}{N(N-1)} = \frac{\Sigma_{j>i}^N (Y_i - Y_j)^2}{N(N-1)} = \sigma^2 \end{aligned}$$

(optional) se of the average

$$\begin{aligned} \text{Var } \bar{Y} &= \text{Var } \left(\frac{1}{n} \Sigma^N S_i Y_i\right) = \\ &= \frac{1}{n^2} \left(\Sigma^N Y_i^2 \text{Var } S_i + 2\Sigma_{j>i}^N Y_i \, y_j \text{ Cov } S_i, S_j\right) \\ \text{E } S_i S_j &= P[S_i = 1, S_j = 1] = ES_{ij} = \frac{n(n-1)}{N(N-1)} \\ \text{Cov } S_i, S_j &= \text{E } S_i S_j - (\text{E } S_i)(\text{E } S_j) = \frac{n(n-1)}{N(N-1)} - \left(\frac{n}{N}\right)^2 \\ &= \frac{-n(N-n)/N}{N(N-1)} \\ \text{Var } \bar{Y} &= \frac{1}{n^2} \frac{n}{N} \frac{N-n}{N} \left(\Sigma^N Y_i^2 - \frac{1}{N-1} \Sigma_{j>i}^N Y_i \, Y_j\right) \end{aligned}$$

(optional) se of the average

This can be simplified by recognizing

$$\begin{aligned} \Sigma^N (Y_i - \bar{Y})^2 &= \Sigma^N Y_i^2 - \frac{(\Sigma^N Y_i)^2}{N} \\ &= \frac{N-1}{N} \left(\Sigma^N Y_i^2 - \frac{1}{N-1} \Sigma_{j>i}^N Y_i \, Y_j\right) \\ \text{Var } \bar{Y} &= \frac{1}{n} \frac{N-n}{N} \frac{\Sigma(Y_i - \bar{Y})^2}{N-1} \\ &= \frac{\sigma^2}{n} \frac{N-n}{N} \end{aligned}$$

- Thompson, *Sampling*, is a good book on all this

Summary of sampling spatial data

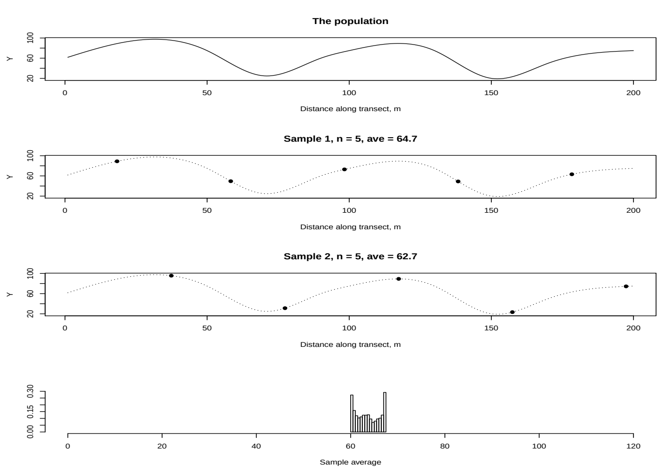
- Notice what was not assumed above:
  - no distribution (no normality)
  - no equal variances
  - no assumption of relationships between neighbors
  - just each obs equally likely to be sampled
  - and each pair equally likely to be sampled
- All the properties of estimators in a simple random sample follow from the random selection of elements from the population.
- In particular, constant joint inclusion probability gets you a valid estimate of the standard error

Spatial correlation and the SRS

- Another way of thinking about spatial correlation and a SRS:
  - The selection of units 1,2,3,4,5 is just as likely as any other sample
  - Can randomly permute the population, no change to properties of the estimators
  - But after permutation, no relationship among neighbors, no spatial correlation
- Having just said all this, there may be better estimators (e.g. of the population mean),
- Better in the sense of having a smaller standard error than the SRS estimator

Systematic spatial samples

- Simple random samples are not commonly used for spatial data
- Systematic sampling is much more common
- Put down a long meter tape and sample (soil, plants, ...) every 10m.
- Or sample at a grid of points, separated by 10m EW and 10m NS
- Best is a random start systematic sample
  - Starting point is randomly chosen, then every X m
- $n = 5$ points on our 200m = 2000 unit transect
  - 200m/5 = 40m between points
  - randomly choose starting point between 0.1m and 40.0m
  - e.g. start at 10.5m. Sample at 10.5m, 50.5m, 90.5m, 130.5, 170.5m



Systematic spatial samples

Statistical properties of systematic sampling

- Because randomness only at the start, only $N/n = 400$ unique samples
- P[unit i is sampled] is same for all units
- so $\bar{Y}$ is unbiased
- Joint inclusion probability, P[units i and j are in the sample], not the same for all pairs
  - 1/400 if i and j separated by multiple of 40.0m
    40m is the spacing of samples along transect
  - 0 if not multiple of 40m apart
- which means big problems estimating Var Y and especially Var $\bar{Y}$.

Systematic sampling on this transect

- Population quantities: $\mu = 63.59$, $\sigma^2 = 551.41$, $\sigma = 23.48$
- Systematic sample: $\bar{Y} = 63.59$, $\text{E}s^2 = 679.33$ (23% larger than σ^2)
- Biggest change: Var $\bar{Y} = 6.13$, much much smaller than $\sigma^2/n \approx 110$.
- So small because for $n = 5$, each systematic sample includes some "high" places and some "low" places.
- Very dependent on the population under study and the relationship between it and the sample
- Can't make generalizations about E s^2: can be "too small" or "too large".
- Traditional example: ag field with high and low places because of plowing. Real problem when sample locations line up with plow lines.
- Worse, don't even know about the problem from the sample information alone.

GrTS sampling

- Systematic sampling has some desirable features
  - Spreads points out.
    - SRS could sample all 5 points between 100m and 110m on our transect
    - Systematic can not.
    - Sample points never "too close" to each other
    - No part of the population "too far" from a sample point
  - Maximizes information when nearby observations are correlated
    - pair of highly correlated (nearby) points has less information
    - well-space points closer to independent
- and some issues
  - difficult to estimate se
    - joint-inclusion probability = 0 for many pairs
- Solutions include
  - multiple systematic samples: analyze as a cluster sample
  - GrTS sampling
- GrTS: Generalized Random Tesselation Stratified Design
  - Stevens, D.L. Jr. and Olsen, A.R. JAgBiolEnvStats 4:415-428 (1999), Environmetrics 14:593-610 (2003), JAmStatAssoc 99:262-278 (2004)

GrTS sampling

- true probability design
  - inclusion and joint inclusion probability are known
  - both > 0, so valid estimate of mean/total and its se
- approximately spatially balanced
  - points spread out, like a systematic sample
- Plus: subsets $L_1 \cdots L_m$, $m < n$ are also spatially balanced
- Common problem with systematic sampling
  - Plan to take $n = 20$ samples from $N = 2000$.
  - Sample L_5, L_{105}, L_{205}, $\cdots$ L_{1505} then a storm blows in
  - Subset is not spatially balanced.
- Useful for monitoring program design
  - Have funding for 20 locations. Draw sample of 50 locations. Sample first 20. If get more $ in the future, add locations from the list of 50.
  - Rotating panel: two types of monitoring locations.
    - Permanent sites: sampled every year
    - Rotating sites: 5 groups, one group sampled each year
    - Denser spatial coverage AND ability to detect sudden change

Design- and model-based inference

- The theory a few slides ago illustrated design-based inference
  - Population values are fixed,
  - the random variables are whether or not unit i included in the sample
- The alternative is to presume a model for the population of values
- e.g. $Y_i \overset{iid}{\sim} N(\mu, \sigma^2)$
- iid: Independent, identically distributed
- If you believe this model, then 3 equally valid samples:
  - 5 randomly chosen units
  - 5 systematically sampled units
  - Y_1, Y_2, Y_3, Y_4, Y_5 (1st five values in the population)
- Validity of inferences depends on validity of the model
- Most statistical methods rely on model-based inference
- Hence so much emphasis on diagnostics to assess assumptions

Sampling without a design

What if you have a happenstance collection of samples?

- No list of items in the population (actual or hypothetical)
- No probability-based selection of sample
- But, no deliberate attempt to select samples with certain properties
  - Example of a deliberate attempt
  - Pigs: average litter size ca 10 piglets / sow
  - Can't measure all, choose 2 largest (by eye) and 2 smallest (by eye)
  - Reasonable estimate of mean, overestimate variance
  - Can get valid estimates using Ranked Set Sampling

Happenstance samples

What can you do?

- Many opinions
- Mine: Is it reasonable to treat sample as if SRS or some other random sample?
- Depends on non-statistical information
- Two examples:
  - Average annual precipitation in continental US
  - Average temperature change (1815 - 2015) in cont. US

Precipitation

Precipitation

- Could calculate average of all ca. 5000 stations
- Should you?
  - Probably not: a particular 0.1 km^2 more likely to be sampled in Midwest / Eastern US
  - Data should not be considered equal probability sample
- Could tessellate the US: e.g., Voronoi = Dirichlet tessellation
  - Polygon i outlines the area closer to point i than any other point.
  - Will be larger in desert areas (precip. stations further apart) than Midwest / Eastern US
- Then consider sample location as a random sample of one location within each polygon
- P[location i in sample] = $p_i \propto 1$ / area of the polygon.
- unequal probability sample. $\hat{\mu} = \Sigma Y_i / p_i$
- result is an area-weighted average.

Temperature change

- Could calculate average of all long-term temperature records
- Should you?
  - Many issues, I'm sure I only know some.
  - More than area sampling issues
  - Precip. analysis assumed that sampled locations are not systematically different from unsampled areas.
  - Most (all?) long term temperature records in cities.
  - Urban heat island effects: cities may systematically differ from rural areas.
- Concept for both: I'm the wrong person to decide whether a happenstance sample provides useful information about the larger population.

Model based approach for happenstance samples

- Alternative: abandon design-based inference. Assume a model.
- No statistical issues, except
  - Validity of inference assumes that model is correct
  - May be hard to justify
- Especially because the population being sampled may not be clearly defined
- Temperature change
  - assume data are a equiprobable sample from some population
  - not clear exactly what that pop. is, but it has a μ.
  - and $\bar{Y}$ estimates μ. (because equiprobable assumption)
  - Not clear that you care about μ

Summary of sampling

- Statistical inference for samples justified by the sample design
  - Design: how sample units were selected
- SRS: valid even if spatial correlation
  - math shown only to give you a flavor for how results can be derived
  - usual estimators are valid but there may be better ones
- When problem is important, spend time thinking about the sampling design
- Or justified by assuming a model
  - Conclusions appropriate when model assumptions are appropriate.

Summary of sampling - 2

- Larger concept: want to estimate or predict some quantity
  - parameter for a population, value at a location
  - more than one way to convert sample values to an estimate / prediction
  - to compare methods:
    - evaluate what happens when sampling is repeated
    - Bias: on average, are we correct?
    - Precision: how variable is the estimate? quantify using se.