# Design and Analysis of Sample Surveys

Andrew Gelman
Department of Statistics and Department of Political Science
Columbia University

Class 1b: Statistical inference and linear regression

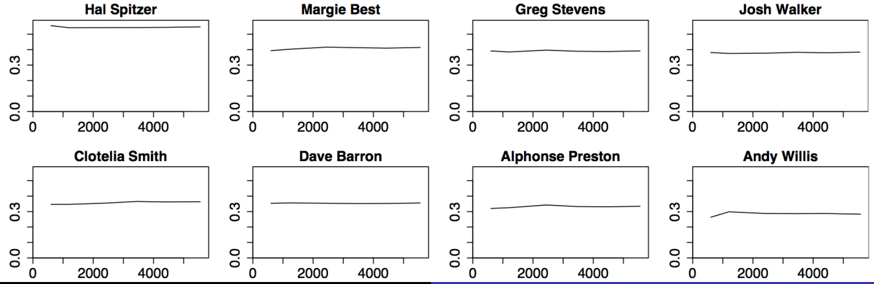
# Statistical inference

- Estimates of proportions
- Estimates and standard errors for continuous parameters
- $1/\sqrt{n}$
- Finite-population correction: $\sqrt{1/n - 1/N}$

# 55,000 residents desperately need your help!

| | | | | | | |
|---|---|---|---|---|---|---|
| Clotelia Smith | 208 | 416 | 867 | 1259 | 1610 | 2020 |
| Earl Coppin | 55 | 106 | 215 | 313 | 401 | 505 |
| Clarissa Montes | 133 | 250 | 505 | 716 | 902 | 1129 |
| ... | ... | ... | ... | ... | ... | ... |

Figure 2.7 *Subset of results from the cooperative board election, with votes for each candidate (names altered for anonymity) tallied after 600, 1200, 2444, 3444, 4444, and 5553 votes. These data were viewed as suspicious because the proportion of votes for each candidate barely changed as the vote counting went on. (There were 27 candidates in total, and each voter was allowed to choose 6 candidates.)*



# Disjoint (instead of cumulative) vote proportions

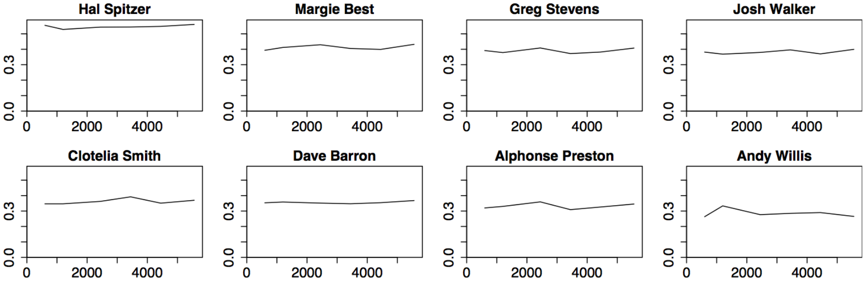



Figure 2.9 *Proportion of votes received by each of the 8 leading candidates in the cooperative board election, at each disjoint stage of voting: the first 600 votes, the next 600, the next 1244, then next 1000, then next 1000, and the final 1109, with the total representing all 5553 votes. The plots here and in Figure 2.8 have been put on a common scale which allows easy comparison of candidates, although at the cost of making it difficult to see details in the individual time series.*

# Comparing to $\sqrt{p(1-p)/n}$

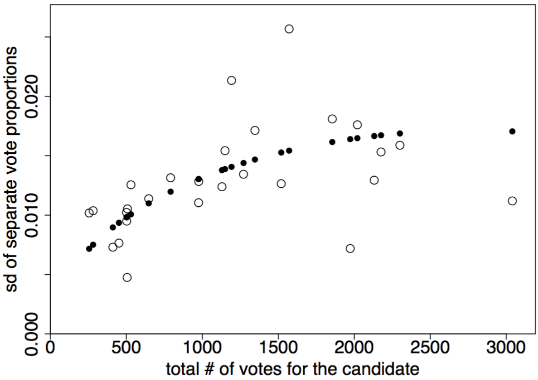



Figure 2.10 *The open circles show, for each of the 27 candidates in the cooperative board election, the standard deviation of the proportions of the vote received by the candidate in the first 600, next 600, next 1244, ..., and the final 1109 votes, plotted versus the total number of votes received by the candidate. The solid dots show the expected standard deviation of the separate vote proportions for each candidate, based on the binomial model*

# Comparisons

- s.e. $= \sqrt{\text{s.e.}_1^2 + \text{s.e.}_2^2}$
- Example: a survey includes 900 U.S.-born and 100 foreign-born adults
- Attitudes on immigration reform: native-born are 60/40 opposed, foreign-born are 80/20 in support
- Estimate and standard error for the comparison?
- General formula:
  - Estimate is $1 \cdot \hat{\theta}_1 + (-1)\hat{\theta}_2$
  - s.e. $= \sqrt{(1)^2\text{s.e.}_1^2 + (-1)^2\text{s.e.}_2^2}$

# Weighted averages

- $y_{\mathrm{w.avg}} = 0.2y_1 + 0.3y_2 + 0.5y_3$
- $\mathrm{sd}(y_{\mathrm{w.avg}}) = \sqrt{0.2^2\mathrm{se}_1^2 + 0.3^2\mathrm{se}_2^2 + 0.5^2\mathrm{se}_3^2}$
- Example: survey with 3 strata:
  - In stratum 1, 75% Yes responses out of 200 respondents
  - In stratum 2, 80% Yes out of 300
  - In stratum 3, 90% Yes out of 400
  - Weighted average: $0.2 \cdot 0.75 + 0.30 \cdot 0.80 + 0.5 \cdot 0.90 = 0.84$
  - Standard error: ?
- Next: example with numerical responses
  - In stratum 1, avg 2.5, sd 0.9, out of 200 respondents
  - In stratum 2, avg 3.0, sd 0.9, out of 300
  - In stratum 3, avg 4.0, sd 1.3, out of 400
  - Weighted average: ?
  - Standard error: ?

# Linear regression

- Interpreting coefficients
- Building models
- The role of statistical significance

# Predicting earnings from height

```
lm(formula = log.earn ~ height)
            coef.est coef.se
(Intercept)     5.74    0.45
height          0.06    0.01
  n = 1192, k = 2
  residual sd = 0.89, R-Squared = 0.06
```

# Predicting earnings from height and sex

```
lm(formula = log.earn ~ height + male)
            coef.est coef.se
(Intercept)    8.15    0.60
height         0.02    0.01
male           0.42    0.07
  n = 1192, k = 3
  residual sd = 0.88, R-Squared = 0.09
```

# Predicting the yield of mesquite bushes

| | |
|---|---|
| diam1: | diameter of the canopy (the leafy area of the bush) in meters, measured along the longer axis of the bush |
| diam2: | canopy diameter measured along the shorter axis |
| canopy.height: | height of the canopy |
| total.height: | total height of the bush |
| density: | plant unit density (# of primary stems per plant unit) |
| group: | group of measurements (0 for the first group, 1 for the second group) |

# Linear model

```
               coef.est coef.se
(Intercept)       -729     147
diam1              190     113
diam2              371     124
canopy.height      356     210
total.height      -102     186
density            131      34
group             -363     100
  n = 46, k = 7
  residual sd = 269, R-Squared = 0.85
```

# Prediction on log scale

```
                      coef.est coef.se
(Intercept)               5.35    0.17
log(diam1)                0.39    0.28
log(diam2)                1.15    0.21
log(canopy.height)        0.37    0.28
log(total.height)         0.39    0.31
log(density)              0.11    0.12
group                    -0.58    0.13
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

# Linear transformation

$$\text{canopy.volume} = \text{diam1} \cdot \text{diam2} \cdot \text{canopy.height}.$$

```
                    coef.est coef.se
(Intercept)             5.17    0.08
log(canopy.volume)      0.72    0.05
  n = 46, k = 2
  residual sd = 0.41, R-Squared = 0.80
```

# More linear transformations

$$\begin{aligned}\text{canopy.area} &= \text{diam1} \cdot \text{diam2} \\ \text{canopy.shape} &= \text{diam1}/\text{diam2}.\end{aligned}$$

```
                    coef.est coef.se
(Intercept)             5.35    0.17
log(canopy.volume)      0.37    0.28
log(canopy.area)        0.40    0.29
log(canopy.shape)      -0.38    0.23
log(total.height)       0.39    0.31
log(density)            0.11    0.12
group                  -0.58    0.13
  n = 46, k = 7
  residual sd = 0.33, R-Squared = 0.89
```

# Stop here?

```
                     coef.est coef.se
(Intercept)              5.31    0.16
log(canopy.volume)       0.38    0.28
log(canopy.area)         0.41    0.29
log(canopy.shape)       -0.32    0.22
log(total.height)        0.42    0.31
group                   -0.54    0.12
  n = 46, k = 6
  residual sd = 0.33, R-Squared = 0.88
```

# Summary of linear regression concepts

- Challenge in interpreting coefficients even in simple models
- Logarithmic transformations to get multiplicative effects
- Why linear transformations are important, even though a naive reading of statistical theory would suggest otherwise

# Homework due beginning of class 3a

- Sample size calculation
- Linear regression in R