

Introduction to Statistics

Global MECOR Course
Kenya, 2011

Getting to Know Your Data

Most good data analysts spend the majority of their time cleaning data and only a relatively small percentage doing formal statistical analyses.

Getting to Know Your Data

Things to look for

- impossible or improbable values
- outliers
- inconsistent or illogical patterns
- missing data, either real or due to skip patterns
- text (i.e., alphanumeric) variables

Getting to Know Your Data

Strategies for exploring your data

- simple frequencies of all categorical variables
- univariate stats (mean, SD, percentiles, minimum and maximum values) for all continuous variables
- selected crosstabs, especially for nested questions (i.e., if “yes” to Q1, then ask Q2)
- listings of selected variables

Getting to Know Your Data

Sample frequency table

Consider the following frequency table for the # asthma hospitalizations in the past year.

| # hospitalizations | N | % | Cum% |
|--------------------|-----|------|------|
| 0 | 71 | 66.4 | 66.4 |
| 1 | 19 | 17.8 | 84.1 |
| 2 | 12 | 11.2 | 95.3 |
| 3 | 3 | 2.8 | 98.1 |
| 5 | 1 | 0.9 | 99.1 |
| 10 | 1 | 0.9 | 100 |
| | 107 | 100 | 100 |

- Are the “5” and “10” values real?
- How might you analyze such data?

Getting to Know Your Data

Sample univariate stats

| arterial pH | | | | |
|-------------|------|----------|-------------|-----------|
| Percentiles | | Smallest | | |
| 1% | 4.42 | 4.42 | | |
| 5% | 7.38 | 7.35 | | |
| 10% | 7.39 | 7.35 | Obs | 71 |
| 25% | 7.41 | 7.38 | Sum of Wgt. | 71 |
| 50% | 7.42 | | Mean | 7.383099 |
| | | Largest | Std. Dev. | .3581324 |
| 75% | 7.44 | 7.49 | | |
| 90% | 7.47 | 7.5 | Variance | .1282588 |
| 95% | 7.49 | 7.51 | Skewness | -8.143764 |
| 99% | 7.52 | 7.52 | Kurtosis | 67.90171 |

- Does anything strike you as peculiar or suspect with this variable?
- The 4.42 was a data entry error. Should be 7.62

Getting to Know Your Data

Dealing with inconsistent response patterns

Consider the following table.

| | | Current smoker? (Q5) | | | total |
|-----------------------|---------|----------------------|----|---------|-------|
| | | yes | no | missing | |
| Ever smoker? (Q5c) | yes | 1 | 21 | 0 | 22 |
| | no | 2 | 16 | 0 | 18 |
| | missing | 53 | 7 | 0 | 60 |
| total | | 56 | 44 | 0 | 100 |

- How might we resolve the 3 people who answered both questions?
- What about the 7 folks who skipped Q5c but shouldn't have?

Getting to Know Your Data

Assigning codes to missing data for logical skips

Consider the following two questions:

| | | |
|--|------------------|------------------|
| 3. Has a doctor ever told you that you have asthma? | No(0):__ | go to Q4 |
| | Yes(1):__ | go to Q3a |

If yes to Q3,

| | |
|---|------------------|
| 3a. Do you have symptoms of cough, shortness of breath, or wheezing? | No(0):__ |
| | Yes(1):__ |

- Q3a will be skipped, and hence be missing, for everyone who answers “no” to Q3.
- Is there a logical value to assign in this case?
- What are merits of assigning “0” (no) vs. NA?

Getting to Know Your Data

Assigning codes to missing data for logical skips

How would you code the following questions?

- age started smoking for those who reported being never smokers in the stem question
- number of asthma hospitalizations in the last year in those who reported never being hospitalized for asthma in the stem question

Getting to Know Your Data

Listing data to check recodes

Look at data to make sure vars were computed correctly

no_b3 should be 1(yes) if no3 \geq 22 & 0(no) else ... (no missing data)

Y2 (outcome at t2) should equal prst3 (status in two weeks prior to t3)

```
list no3 no_b3 prst3 Y2 in 1/20
```

| | no3 | no_b3 | prst3 | Y2 |
|-----|-------|-------|-------|-------|
| 1. | 29.4 | 1=yes | 0=no | 0=no |
| 2. | 17.2 | 0=no | 1=yes | 1=yes |
| 3. | 47.2 | 1=yes | 1=yes | 1=yes |
| 4. | 46.7 | 1=yes | 0=no | 0=no |
| 5. | 19.5 | 0=no | 0=no | 0=no |
| 6. | 12.1 | 0=no | 0=no | 0=no |
| 7. | 17.7 | 0=no | 0=no | 0=no |
| 8. | 36.8 | 1=yes | 1=yes | 1=yes |
| 9. | 71.9 | 1=yes | 0=no | 0=no |
| 10. | 31.4 | 1=yes | 1=yes | 1=yes |
| 11. | 52.9 | 1=yes | 1=yes | 1=yes |
| 12. | 20.7 | 0=no | 0=no | 0=no |
| 13. | 71.85 | 1=yes | 1=yes | 1=yes |
| 14. | 32.6 | 1=yes | 1=yes | 1=yes |
| 15. | 21.8 | 0=no | 0=no | 0=no |
| 16. | 55 | 1=yes | 1=yes | 1=yes |
| 17. | 53 | 1=yes | 0=no | 0=no |
| 18. | 28.7 | 1=yes | 0=no | 0=no |
| 19. | 21.3 | 0=no | 1=yes | 1=yes |
| 20. | 45.1 | 1=yes | 0=no | 0=no |

Getting to Know Your Data

The Bottom Line:

Garbage In = Garbage Out !

Spending the time getting to know and understand your data will pay off in the long run.



Statistics

Inside the Black Box



Inside the Black Box:

Maximum Likelihood Estimation

Most people have heard of the normal distribution. When we say that some variable is normally distributed with mean, μ , and variance, σ^2 , we are tacitly assuming that we can write an equation describing the probability (or likelihood) of the observed data as a function of μ and σ^2 .

The values of μ and σ^2 that maximize the probability are termed “**maximum likelihood estimates**” (MLEs).

Inside the Black Box:

Regression modeling and maximum likelihood

Whenever you fit a regression model, you are asking the computer to generate maximum likelihood estimates. However rather than simply estimate a single overall mean, μ , we typically want to describe the mean in terms of other explanatory variables. For example,

$$\text{mean FEV}_1 = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$$

The coefficients in this model (the β s) are also MLEs!

Inside the Black Box:

Properties of regression coefficients

Regression coefficients have two very desirable properties for statisticians:

- If the source data are normally distributed, then the regression coefficients will be normally distributed.
- Even if the source data are not normally distributed, the regression coefficients derived from such data will be \approx normally distributed for large enough sample sizes.

We use these properties to test specific hypotheses of interest (e.g., $H_0: \beta_1=0$).

Inside the Black Box:

Some common distributions & regression models

In addition to the normal distribution and standard linear regression, other common distributions and regression models used in the medical literature are:

- the **binomial distribution**, which forms the basis for **logistic regression** and is used to analyze **binary** (yes/no) data
- the **Poisson distribution**, useful for modeling **rates of occurrence** via **Poisson regression**, and
- the **Cox proportional hazards model**, used to analyze **time to event data**.

Inside the Black Box:

Building and interpreting regression models

Each distribution gives rise to an equation that relates a basic parameter of the model to a collection of predictor variables.

e.g.,

normal: $\mu = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$

binomial: $\ln[P/(1-P)] = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$

Cox: $\ln[\lambda(t)] = \ln[\lambda_0(t)] + \beta_1 \text{Age} + \beta_2 \text{Height}$
 $\approx \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Height}$

Inside the Black Box:

Building and interpreting regression models

Adjusting for *potential confounders*:

- To adjust for potential confounders, we add main effect terms to our model

unadjusted

- $\ln(\text{odds of death}) = \beta_0 + \beta_1 \text{Male}$

adjusted for ever smoking status

- $\ln(\text{odds of death}) = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{EvSmk}$

$\ln(\text{OR}_{\text{Mvs.F}})$ is the same for males and females (β_1)

Inside the Black Box:

Building and interpreting regression models

Adjusting for *effect modifiers*:

- To adjust for effect modifiers, we add interaction terms to our model

$$\ln(\text{odds of death}) = \beta_0 + \beta_1 \text{Male} + \beta_2 \text{EvSmk} + \beta_3 \text{Male} * \text{EvSmk}$$

- To simplify interpretation of the model, we also include a main effect term for the effect modifier
- What does this model say about OR for M vs. F?

$$\ln(\text{odds of death}) = \beta_0 + \beta_1 \text{Male} \quad (\text{never smokers})$$

$$\ln(\text{odds of death}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Male} \quad (\text{ever smokers})$$

Inside the Black Box:

Building and interpreting regression models

Adjusting for *effect modifiers*:

- $\ln(\text{odds}) = \beta_0 + \beta_1 \text{Male}$ (never smokers)
- $\ln(\text{odds}) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Male}$ (ever smokers)

Solve for OR of death in M vs. F:

- $\ln(\text{OR}_{\text{Mvs.F}}) = \beta_1$ (never smokers)
- $\ln(\text{OR}_{\text{Mvs.F}}) = \beta_1 + \beta_3$ (ever smokers)

So in effect modification, the effect of our primary exposure variable on the outcome of interest varies according to the level of the effect modifier.

Inside the Black Box:

Building and interpreting regression models

Regardless of the underlying distribution, the principles remain the same.

- If we are interested in the effect of gender on some outcome (FEV1, death, rate of hospitalizations) and we want to adjust for some potential **confounder**, we add that confounder as a **main effect term** to our model.
- If we want to see if the **effect** of gender on our outcome **differs** by smoking status or race or some other factor, we add an **interaction** of gender with that variable to our model.