

Lecture 12

Statistics

Ryan McWay[†]

[†]*Applied Economics,
University of Minnesota*

Mathematics Review Course, Summer 2023
University of Minnesota
August 22nd, 2023

LAST LECTURE REVIEW

- Probability:
 - Probability Limits
 - Independence
 - Law of Total Probability
 - Conditional Probability
 - Cumulative Distribution Function
 - Probability Distribution Function
 - Joint & Marginal Distributions
 - Gaussian (Normal) Distribution
 - Bayes Rules
 - Moments of a Distribution
 - Covariance & Correlation

REVIEW ASSIGNMENT

1. Problem Set 11 solutions are available on Github.
2. Any issues or problems **You** would like to discuss?

Statistics

STATS: GARBAGE IN \rightarrow GARBAGE OUT

PRECISE NUMBER + PRECISE NUMBER = SLIGHTLY LESS PRECISE NUMBER

PRECISE NUMBER \times PRECISE NUMBER = SLIGHTLY LESS PRECISE NUMBER

PRECISE
NUMBER + GARBAGE = GARBAGE

PRECISE
NUMBER

x GARBAGE = GARBAGE

$$\sqrt{\text{GARBAGE}} = \text{LESS BAD GARBAGE}$$

$(\text{GARBAGE})^2 = \text{WORSE GARBAGE}$

$$\frac{1}{N} \sum (N \text{ PIECES OF STATISTICALLY INDEPENDENT GARBAGE}) = \text{BETTER GARBAGE}$$

(PRECISE) GARBAGE = MUCH WORSE GARBAGE

GARBAGE - GARBAGE = MUCH WORSE
GARBAGE

$\frac{\text{PRECISE NUMBER}}{\text{GARBAGE} - \text{GARBAGE}} = \text{MUCH WORSE GARBAGE, POSSIBLE DIVISION BY ZERO}$

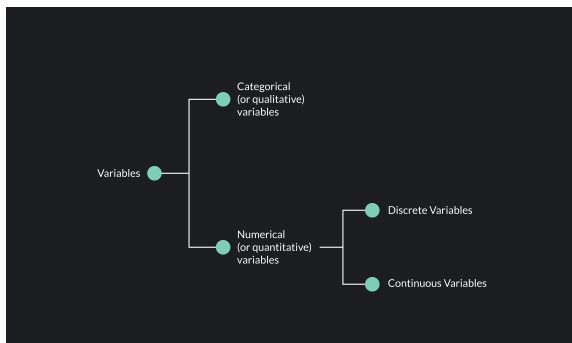
GARBAGE \times 0 = PRECISE
NUMBER

2. RANDOM VARIABLES

- ▶ Experiment: Procedure that can be infinitely repeated and has well-defined set of outcomes. But, the actual outcome may be of unknown certainty.
- ▶ Random variable: $x \in X$ is a numerical value outcome in a set of possible outcomes where the outcome is determined by an experiment.
- ▶ Binary (Bernoulli) random variables: Takes values 0 or 1 such that $Pr(X = 1) = \theta : 0 \leq \theta \leq 1$.

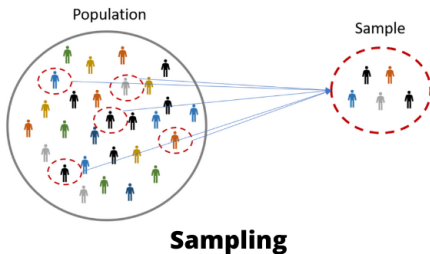
3. DISCRETE & CONTINUOUS VARIABLES

- ▶ Discrete Variables: Random variable that has a finite or countably infinite domain of values.
- ▶ Continuous Variables: Random variable with infinite domain of values.
 - ▶ Can take on any real value with $Pr(X) > 0$.

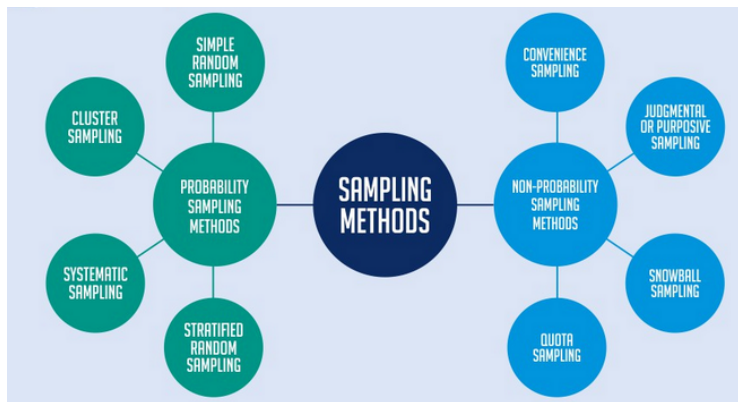


6. SAMPLING

- ▶ Sampling pulls units from the population distribution as the data used to estimate population parameters.
- ▶ Random Sampling
 - ▶ A sample of independently drawn units.
 - ▶ Independence means that no individual is more likely to be sampled than another from the population.



6. SAMPLING



7. DATA TYPES

- ▶ The data you gather may come in many forms – qualitative and quantitative.
- ▶ Most of the data you will work with are samples formatted into tabular arrays (viz., matrix).
- ▶ Types:
 - ▶ Float: Fractions
 - ▶ Boolean: True/False
 - ▶ Integer: \mathbb{Z}
 - ▶ Categorical: Qualitative information
 - ▶ Character: Letter or special character
 - ▶ String: Combination of characters
 - ▶ Date: Formatted string or integer
 - ▶ Null: Empty Set
 - ▶ N/A: Missing value

8. DATA STRUCTURE

- ▶ The data generating process (DGP)
 - ▶ The way that the data is constructed in the real world.
 - ▶ Typically unknown – often an assumption.
- ▶ Independently Identically Distributed (IID)
 - ▶ Each random variable has the same probability distribution as the others and are mutually independent of one another.
 - ▶ The sample is ‘representative’ of the population.
 - ▶ Identically distributed: Distribution does not fluctuate by sample.
 - ▶ Independent: Sample items (outcomes) are independent events

Probabilities of sequences of independent and identically distributed random events

N = 1	N = 2	N = 3
$p(\text{blue}) = \frac{1}{2}$	$p(\text{blue blue}) = \frac{1}{4}$	$p(\text{blue blue blue}) = \frac{1}{8}$
$p(\text{grey}) = \frac{1}{4}$	$p(\text{grey green}) = \frac{1}{32}$	$p(\text{grey green grey}) = \frac{1}{128}$
$p(\text{yellow}) = \frac{1}{8}$	$p(\text{blue yellow}) = \frac{1}{16}$	$p(\text{blue yellow yellow}) = \frac{1}{128}$
$p(\text{green}) = \frac{1}{8}$	\vdots	\vdots

8. DATA STRUCTURE

- ▶ Cross-sectional
 - ▶ A sample of individuals (units) at a given point in time (period) with many variables (characteristics).
- ▶ Time Series
 - ▶ Potentially several units on a single variable over many periods.
- ▶ Pool Cross-section
 - ▶ A series of cross-sections with some overlapping units at different time periods.
- ▶ Panel Datasets
 - ▶ A time series of many periods for many variables across many units.

9. RANDOMIZATION

- ▶ A random process is a random sequence that does not follow a deterministic process.
- ▶ Random Sample: Individuals have a known probability of sampling from the population.
- ▶ Ensure the data is I.I.D. and that the estimates are representative for the parameters in the population.

10. ESTIMATE, ESTIMATOR, & ESTIMAND

- ▶ Estimate: Approximation of the population parameter.
- ▶ Estimator: The function or algorithm doing the estimation.
- ▶ Estimand: The parameter in the population you aim to estimate.

ESTIMAND
What you seek



E.g. The true difference in Y
due to exposure

ESTIMATOR
How you will get there

Method

1. Preheat your oven to 180°C / 350°F / Gas Mark 5. Grease and line the base of 2 cake tins, one 8 inch/20cm and one 6 inch/15cm.
2. Cream together the butter and caster sugar until light and fluffy.
3. Add the eggs one at a time with a spoonful of flour and blend in well.
4. Sift in the flour and baking powder and gently fold in. Finally add the milk and mix until you have a smooth batter.
5. Pour 1/3 of the batter into the small tin and 2/3 into the large tin.
6. Bake on the same shelf in the preheated oven, the smaller tin at the front.
7. Check the smaller cake after 20 minutes. When it is cooked remove from the oven, leaving the larger one still baking. The larger cake should be done by 30 minutes.
8. Leave the cakes for 5 minutes in the tins, then turn out onto a rack to cool completely.
9. To make the icing beat together the butter and icing sugar, add the vanilla and then the milk. Whisk the icing hard using an electric stand mixer if you can. Whisk it for 5 minutes and it will become really pale and light.

E.g. Your regression
model

ESTIMATE
What you get



E.g. the estimated difference
in Y from model coefficient

11. PARAMETRIC VS. NON-PARAMETRIC

- ▶ Parametric: Imposes some distribution on the underlying population (e.g., the distribution of the parameters)
- ▶ Non-parametric: Is agnostic to how the distribution of the population might look
- ▶ Parametric estimators rely on asymptotic properties of the population (e.g., Assume to know the underlying distribution in large samples).
- ▶ Non-parametric estimators are driven by distributional properties of the sample.
- ▶ Can often use non-parametric estimators when the moments of the distribution do not converge to a known distribution.

12. EXPECTATIONS

- ▶ Expected Values (Mean): A weighted average of all possible values of X , with weights equal to the probability of each outcome occurring.
- ▶ Discrete:

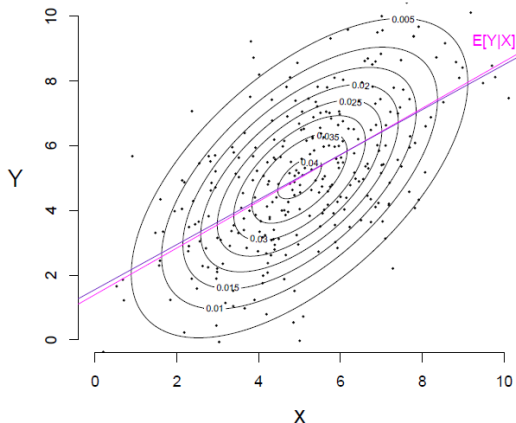
$$\mathbb{E}(X) = \sum_j^k x_j f(x_j) = \mu_x$$

- ▶ Continuous:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

13. CONDITIONAL EXPECTATION FUNCTION

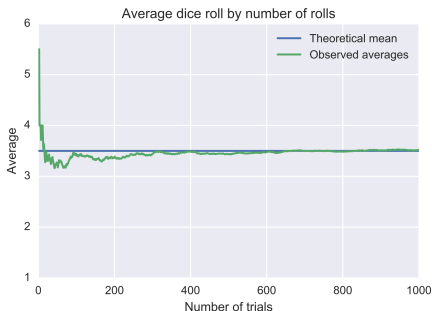
$$\mathbb{E}[Y|X = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x) dy$$



14. LAW OF LARGE NUMBERS

- ▶ The expected value of the distribution converges to the population parameter as the sample becomes larger.
- ▶ If X_i are I.I.D. and $\mathbb{E}[X] < \infty$, then as $n \rightarrow \infty$:

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i \xrightarrow{p} \mathbb{E}[X]$$

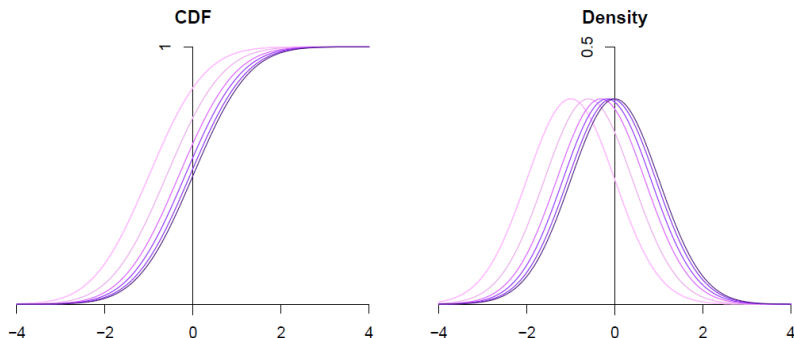


15. CENTRAL LIMIT THEOREM

- ▶ The distribution of the sample approximates a normal distribution as the sample becomes larger – **regardless of the population's underlying distribution.**
- ▶ Let $X_i \in \mathbb{R}^k$ be I.I.D. with $\mathbb{E}[||X_i||^2] < \infty$.
- ▶ Define $\mu = \mathbb{E}[X]$.
- ▶ Define $\Sigma = \mathbb{E}[(X - \mu)(X - \mu)^T]$
- ▶ Then, as $n \rightarrow \infty$:

$$\sqrt{n}(\bar{X} - \mu) \xrightarrow{d} N(0, \Sigma)$$

15. CENTRAL LIMIT THEOREM



16. CONTINUOUS MAPPING THEOREM

- ▶ If a sequence converges, then you can apply a function/transformation to the sequence and it converges to the function of that limit.
- ▶ Idea: $g(Z_n) = g(\text{plim} Z_n)$
- ▶ So: $\hat{\beta} \xrightarrow{p} \beta$
- ▶ If $Z_n \xrightarrow{p} c$ as $n \rightarrow \infty$, and $h(\cdot)$ is continuous at c , then:

$$h(Z_n) \xrightarrow{p} h(c) \text{ as } n \rightarrow \infty$$

17. DELTA METHOD

- ▶ Variance will be approximately normal.
- ▶ Let $\theta \in \mathbb{R}^k$ and $h : \mathbb{R}^k \rightarrow \mathbb{R}$ be C^1 in the neighborhood of θ .
- ▶ Let $h = \nabla h(\theta)$.
- ▶ If $\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} \eta$, then as $n \rightarrow \infty$,

$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} h^T \cdot \eta$$

- ▶ If $\eta \sim N(0, V)$ (by CLT), then





$$\sqrt{n}(h(\hat{\theta}) - h(\theta)) \xrightarrow{d} N(0, h^T V h)$$

18. STANDARDIZING UNITS

- ▶ Commonly referred to as ‘taking a z-score’.
- ▶ De-meaning the random variable and weighting it by the standard error to create a new random variable measured in standard deviations from the mean.
- ▶ When the units are standard deviations, it is now easy to compare many types of outcomes with various ranges and distributions.
- ▶ $\mathbb{E}[Z] = 0$
- ▶ $Var(Z) = 1$

$$Z = \frac{x - \mu_x}{\sigma_x}$$

19. HYPOTHESIS TESTING

	Null Hypothesis is TRUE	Null Hypothesis is FALSE
Reject null hypothesis	 Type I Error (False positive)	 Correct Outcome! (True positive)
Fail to reject null hypothesis	 Correct Outcome! (True negative)	 Type II Error (False negative)

TYPES OF ERRORS

TYPE I ERROR: FALSE POSITIVE

TYPE II ERROR: FALSE NEGATIVE

TYPE III ERROR: TRUE POSITIVE FOR
INCORRECT REASONS

TYPE IV ERROR: TRUE NEGATIVE FOR
INCORRECT REASONS

TYPE V ERROR: INCORRECT RESULT WHICH
LEADS YOU TO A CORRECT
CONCLUSION DUE TO
UNRELATED ERRORS

TYPE VI ERROR: CORRECT RESULT WHICH
YOU INTERPRET WRONG

TYPE VII ERROR: INCORRECT RESULT WHICH
PRODUCES A COOL GRAPH

TYPE VIII ERROR: INCORRECT RESULT WHICH
SPARKS FURTHER RESEARCH
AND THE DEVELOPMENT OF
NEW TOOLS WHICH REVEAL
THE FLAW IN THE ORIGINAL
RESULT WHILE PRODUCING
NOVEL CORRECT RESULTS

TYPE IX ERROR: THE RISE OF SKYWALKER

Probability & statistical significance



20. CAUSATION

- ▶ Considering the ‘counterfactual’ if a treatment D had not occurred.
- ▶ Using randomization to determine the average treatment effect (ATE), since we cannot recover the individual effect.

$$\delta_i = y_i^1 - y_i^0$$

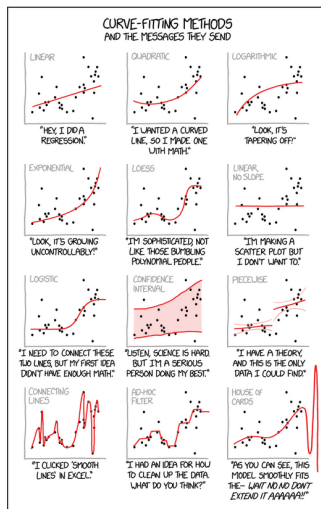
$$Y = DY^1 + (1 - D)Y^0$$

$$\mathbb{E}[\delta] = \mathbb{E}[Y^1] - \mathbb{E}[Y^0]$$

Table 2.1 The Fundamental Problem of Causal Inference

Group	Y^1	Y^0
Treatment group ($D = 1$)	Observable as Y	Counterfactual
Control group ($D = 0$)	Counterfactual	Observable as Y

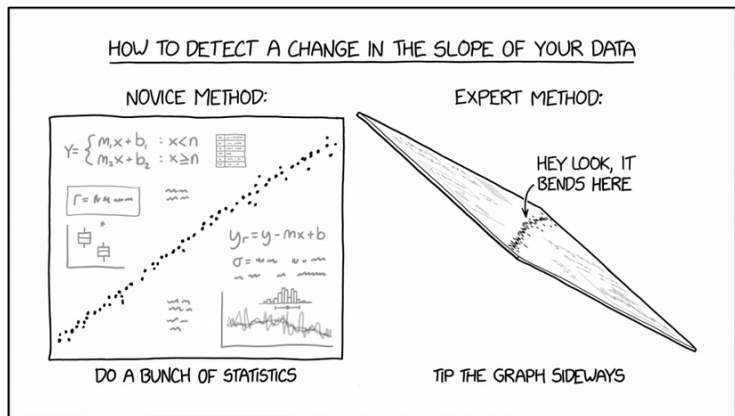
TRY TO NOT OVER-FIT



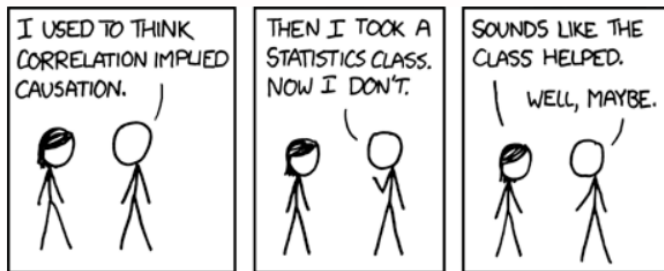
22. DESCRIPTIVE ANALYSIS

- ▶ Rather than determine cause and effect or forecasting the future, we are creating a description of the the current phenomenon.
- ▶ Can we estimate moments of the distribution for important variables that are representative of the population?
- ▶ Can we fit data to a model that approximates the phenomenon? What happens if we change the variables in this model?

ANALYSIS CAN BE HARD



CORRELATION \neq CAUSATION



Review

REVIEW: STATISTICS

- | | | |
|--|--------------------------------------|--------------------------------|
| 1. Population, Parameters, and Distributions | 9. Estimate, Estimator, & Estimand | 15. Continuous Mapping Theorem |
| 2. Random Variables | 10. Parametric vs. Non-parametric | 16. Delta Method |
| 3. Discrete & Continuous Variables | 11. Expectations | 17. Standardizing Units |
| 4. Law of Iterated Expectations | 12. Conditional Expectation Function | 18. Hypothesis Testing |
| 5. Sampling | 13. Law of Large Numbers | 19. Causation |
| 6. Data Types | 14. Central Limit Theorem | 20. Prediction |
| 7. Data Structure | | 21. Descriptive Analysis |
| 8. Randomization | | 22. Inferential Analysis |

