

Basic Principles of Statistical Inference

Kosuke Imai

Department of Politics
Princeton University

POL572 Quantitative Analysis II
Spring 2012

Welcome to POL 572

- Future of social science: Data, data, data ...
- Quantitative Methods Training at Princeton
 - Train a sophisticated user of statistics rather than a blind consumer
 - Rigorous and painful (especially first year)
 - Requires commitments from both instructors and students
- **First Year: Mastering basics**
 - 1 POL 571: Probability Boot Camp
 - 2 **POL 572: Statistics Boot Camp**
- **Second Year: Learning how to do research**
 - 1 POL 573: Statistics Workshop
 - 2 POL 574: Advanced Statistics Workshop
- **Third Year and Beyond: Doing research**
- **Goal:** Learn how to do (good) research and get a (good) job

How Not to Study Statistics

- Learning statistics (or math in general) is **different** from learning substantive topics in political science
- “Skip equations, and try to understand them from the context...”
- “I should be able to get the intuition without knowing the math...”
- “Move on, and try to figure out the next topic...”
- “Do you understand so far? Uhhh... Yes.”
- “I put all of the formulae on flashcards to study for the exam...”
- “I am just going to copy down the formula...”
- “What formula am I supposed to use here?”

What is Statistics?

- Relatively new discipline
- Scientific revolution in the 20th century
- Data and computing revolutions in the 21st century
- The world is stochastic rather than deterministic
- Probability theory used to model stochastic events

- **Statistical inference**: Learning about what we do not observe (parameters) using what we observe (data)
- Without statistics: **wild guess**
- With statistics: **principled guess**
 - 1 assumptions
 - 2 formal properties
 - 3 measure of uncertainty

Three Modes of Statistical Inference

- 1 **Descriptive Inference**: summarizing and exploring data
 - Inferring “ideal points” from rollcall votes
 - Inferring “topics” from texts and speeches
 - Inferring “social networks” from surveys
- 2 **Predictive Inference**: forecasting out-of-sample data points
 - Inferring future state failures from past failures
 - Inferring population average turnout from a sample of voters
 - Inferring individual level behavior from aggregate data
- 3 **Causal Inference**: predicting counterfactuals
 - Inferring the effects of ethnic minority rule on civil war onset
 - Inferring *why* incumbency status affects election outcomes
 - Inferring whether the lack of war among democracies can be attributed to regime types

Statistics for Social Scientists

- Quantitative social science research:
 - ① Find a substantive question
 - ② Construct theory and hypothesis
 - ③ Design an empirical study and collecting data
 - ④ Use statistics to analyze data and test hypothesis
 - ⑤ Report the results
- No study in the social sciences is perfect
- Use best available methods and data, but be aware of limitations
- Many wrong answers but no single right answer
- Credibility of data analysis:

$$\text{Data analysis} = \underbrace{\text{assumption}}_{\text{subjective}} + \underbrace{\text{statistical theory}}_{\text{objective}} + \underbrace{\text{interpretation}}_{\text{subjective}}$$

- Statistical methods are no substitute for a good research design

Sample Surveys

- A large population of size N
 - Finite population: $N < \infty$
 - Super population: $N = \infty$
- A simple random sample of size n
 - Probability sampling: e.g., stratified, cluster, systematic sampling
 - Non-probability sampling: e.g., quota, volunteer, snowball sampling
- The population: X_i for $i = 1, \dots, N$
- Sampling (binary) indicator: Z_1, \dots, Z_N
- **Assumption:** $\sum_{i=1}^N Z_i = n$ and $\Pr(Z_i = 1) = n/N$ for all i
- # of permutations: $\binom{N}{n} = \frac{N!}{n!(N-n)!}$
- Estimand = population mean vs. Estimator = sample mean:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^N Z_i X_i$$

Estimation of Population Mean

- Key idea: Randomness comes from sampling alone
- Unbiasedness (over repeated sampling): $\mathbb{E}(\bar{X}) = \bar{X}$
- Variance of sampling distribution:

$$\mathbb{V}(\bar{X}) = \underbrace{\left(1 - \frac{n}{N}\right)}_{\text{finite population correction}} \frac{S^2}{n}$$

where $S^2 = \sum_{i=1}^N (X_i - \bar{X})^2 / (N - 1)$ is the population variance

- Unbiased estimator of the variance:

$$\hat{\sigma}^2 \equiv \left(1 - \frac{n}{N}\right) \frac{s^2}{n} \quad \text{and} \quad \mathbb{E}(\hat{\sigma}^2) = \mathbb{V}(\bar{X})$$

where $s^2 = \sum_{i=1}^n Z_i (X_i - \bar{x})^2 / (n - 1)$ is the sample variance

- **Plug-in (sample analogue) principle**

Analytical Details of Randomization Inference

- 1 $\mathbb{E}(Z_i) = \mathbb{E}(Z_i^2) = n/N$ and $\mathbb{V}(Z_i) = \mathbb{E}(Z_i^2) - \mathbb{E}(Z_i)^2 = \frac{n}{N} (1 - \frac{n}{N})$
- 2 $\mathbb{E}(Z_i Z_j) = \mathbb{E}(Z_i | Z_j = 1) \Pr(Z_j = 1) = \frac{n(n-1)}{N(N-1)}$ for $i \neq j$ and thus $\text{Cov}(Z_i, Z_j) = \mathbb{E}(Z_i Z_j) - \mathbb{E}(Z_i)\mathbb{E}(Z_j) = -\frac{n}{N(N-1)} (1 - \frac{n}{N})$
- 3 Use these results to derive the expression:

$$\begin{aligned}\mathbb{V}(\bar{X}) &= \frac{1}{n^2} \mathbb{V} \left(\sum_{i=1}^N Z_i X_i \right) \\ &= \frac{1}{n^2} \left\{ \sum_{i=1}^N X_i^2 \mathbb{V}(Z_i) + \sum_{i=1}^N \sum_{j \neq i}^N X_i X_j \text{Cov}(Z_i, Z_j) \right\} \\ &= \frac{1}{n} \left(1 - \frac{n}{N}\right) \underbrace{\frac{1}{N(N-1)} \left\{ N \sum_{i=1}^N X_i^2 - \left(\sum_{i=1}^N X_i \right)^2 \right\}}_{=S^2}\end{aligned}$$

where we used the equality $\sum_{i=1}^N (X_i - \bar{X})^2 = \sum_{i=1}^N X_i^2 - N\bar{X}^2$

Finally, we proceed as follows:

$$\begin{aligned}\mathbb{E} \left\{ \sum_{i=1}^N Z_i (X_i - \bar{x})^2 \right\} &= \mathbb{E} \left[\sum_{i=1}^N Z_i \underbrace{\left\{ (X_i - \bar{X}) + (\bar{X} - \bar{x}) \right\}}_{\text{add \& subtract}}^2 \right] \\ &= \mathbb{E} \left\{ \sum_{i=1}^N Z_i (X_i - \bar{X})^2 - n(\bar{X} - \bar{x})^2 \right\} \\ &= \mathbb{E} \left\{ \sum_{i=1}^N Z_i (X_i - \bar{X})^2 \right\} - n\mathbb{V}(\bar{x}) \\ &= \frac{n(N-1)}{N} S^2 - \left(1 - \frac{n}{N}\right) S^2 \\ &= (n-1)S^2\end{aligned}$$

Thus, $\mathbb{E}(s^2) = S^2$, implying that the sample variance is unbiased for the population variance

Unequal Probability Sampling

- Suppose $\Pr(Z_i = 1) = \pi_i$ differing across units
- We still randomly sample n units from the population of size N where $\sum_{i=1}^N \pi_i = n$
- Oversampling of minorities, difficult-to-reach individuals, etc.
- Inverse probability weighting via the **Horvitz-Thompson estimator**:

$$\tilde{x} = \frac{1}{N} \sum_{i=1}^N \frac{Z_i X_i}{\pi_i}$$

- Unbiasedness: $\mathbb{E}(\tilde{x}) = \bar{X}$
- Variance:

$$\mathbb{V}(\tilde{x}) = \frac{1}{N^2} \left\{ \sum_{i=1}^N (1 - \pi_i) \frac{X_i^2}{\pi_i} + \sum_{i=1}^N \sum_{j \neq i}^N (\pi_{ij} - \pi_i \pi_j) \frac{X_i X_j}{\pi_i \pi_j} \right\}$$

where $\pi_{ij} = \Pr(Z_i = Z_j = 1)$

- What is the unbiased estimator of the variance?

Model-Based vs. Design-Based Approaches

- An infinite population characterized by a probability distribution
 - Nonparametric \mathcal{F}
 - Parametric \mathcal{F}_θ (e.g., $\mathcal{N}(\mu, \sigma^2)$)
- Probability model: a hypothetical population
- A simple random sample of size n : X_1, \dots, X_n
- **Assumption**: X_i is independently and identically distributed (i.i.d.) according to \mathcal{F}
- Estimator = sample mean vs. Estimand = population mean:

$$\hat{\mu} \equiv \frac{1}{n} \sum_{i=1}^n X_i \quad \text{and} \quad \mu \equiv \mathbb{E}(X_i)$$

- Unbiasedness: $\mathbb{E}(\hat{\mu}) = \mu$
- Variance and its unbiased estimator:

$$\mathbb{V}(\hat{\mu}) = \frac{\sigma^2}{n} \quad \text{and} \quad \hat{\sigma}^2 \equiv \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu})^2$$

(Weak) Law of Large Numbers (LLN)

- If $\{X_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2 , then

$$\bar{X}_n \xrightarrow{p} \mu$$

where “ \xrightarrow{p} ” denotes the **convergence in probability**, i.e.,

$$\lim_{n \rightarrow \infty} \Pr(|X_n - \mu| > \epsilon) = 0 \text{ for any } \epsilon > 0$$

- If $X_n \xrightarrow{p} x$, then for any continuous function $f(\cdot)$, we have

$$f(X_n) \xrightarrow{p} f(x)$$

- Implication: Justifies the plug-in (sample analogue) principle

LLN in Action

- In *Journal of Theoretical Biology*,

- ① “Big and Tall Parents have More Sons” (2005)
- ② “Engineers Have More Sons, Nurses Have More Daughters” (2005)
- ③ “Violent Men Have More Sons” (2006)
- ④ “Beautiful Parents Have More Daughters” (2007)



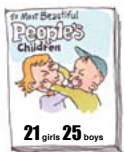
1995



1996



1997



1998



1999



2000

- Law of Averages in action

- ① 1995: 57.1%
- ② 1996: 56.6
- ③ 1997: 51.8
- ④ 1998: 50.6
- ⑤ 1999: 49.3
- ⑥ 2000: 50.0

- No duplicates: 47.7%

- Population frequency: 48.5%

Gelman & Weakliem, *American Scientist*

Central Limit Theorem (CLT)

- If $\{X_i\}_{i=1}^n$ is a sequence of i.i.d. random variables with mean μ and finite variance σ^2 , then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} \mathcal{N}(0, 1)$$

where “ \xrightarrow{d} ” represents the **convergence in distribution**, i.e.,

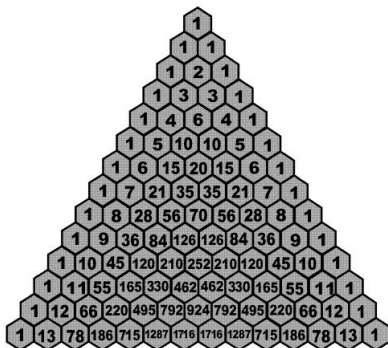
$$\lim_{n \rightarrow \infty} P(X_n \leq x) = P(X \leq x) \text{ for all } x$$

with $P(X \leq x)$ being continuous at every x

- If $X_n \xrightarrow{d} X$, then for any continuous function $f(\cdot)$,

$$f(X_n) \xrightarrow{d} f(X)$$

- Implication: Justifies asymptotic (normal) approximation



Pascal's Triangle

- n^{th} row and k^{th} column = $\binom{n-1}{k-1}$ = # of ways to get there
- Binomial distribution: $\Pr(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$
- Sir Francis Galton's Quincunx

Asymptotic Properties of the Sample Mean

- Infinite (super) population $n \rightarrow \infty$
- LLN implies **consistency**:

$$\hat{\mu} = \bar{X}_n \xrightarrow{p} \mu$$

- CLT implies **asymptotic normality**:

$$\begin{aligned} \sqrt{n}(\hat{\mu} - \mu) &\xrightarrow{d} \mathcal{N}(0, \sigma^2) \\ \implies \hat{\mu} &\underset{\text{approx.}}{\sim} \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{in a large sample} \end{aligned}$$

But, σ is unknown

- **Standard error**: estimated standard deviation of sampling distribution

$$\text{s.e.} = \frac{\hat{\sigma}}{\sqrt{n}}$$

where $\hat{\sigma}^2$ is unbiased (shown before) and consistent for σ^2 (LLN)

Asymptotic Confidence Intervals

- Putting together, we have:

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\hat{\sigma}} = \underbrace{\frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}}}_{z\text{-score}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- We used the **Slutzky Theorem**: If $X_n \xrightarrow{p} x$ and $Y_n \xrightarrow{d} Y$, then $X_n + Y_n \xrightarrow{d} x + Y$ and $X_n Y_n \xrightarrow{d} xY$
- This gives 95% asymptotic confidence interval:

$$\Pr\left(-1.96 \leq \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \leq 1.96\right) \xrightarrow{p} 0.95$$
$$\implies \Pr\left(\hat{\mu} - 1.96 \times \hat{\sigma}/\sqrt{n} \leq \mu \leq \hat{\mu} + 1.96 \times \hat{\sigma}/\sqrt{n}\right) \xrightarrow{p} 0.95$$

- $(1 - \alpha) \times 100\%$ asymptotic confidence interval (symmetric and balanced):

$$CI_{1-\alpha} = [\hat{\mu} - z_{\alpha/2} \times \text{s.e.}, \hat{\mu} + z_{\alpha/2} \times \text{s.e.}]$$

where s.e. represents the standard error

- **Critical value:** $\Pr(Z > z_{\alpha/2}) = \Phi(-z_{\alpha/2}) = \alpha/2$ where $Z \sim \mathcal{N}(0, 1)$
 - ① $\alpha = 0.01$ gives $z_{\alpha/2} = 2.58$
 - ② $\alpha = 0.05$ gives $z_{\alpha/2} = 1.96$
 - ③ $\alpha = 0.10$ gives $z_{\alpha/2} = 1.64$
- Be careful about the interpretation!
 - Confidence intervals are *random*, while the truth is *fixed*
 - Probability that the true value is in a particular confidence interval is either 0 or 1 and not $1 - \alpha$
- Nominal vs. actual coverage probability: $\Pr(\mu \in CI_{1-\alpha}) \xrightarrow{P} 1 - \alpha$

Inference with Normally Distributed Data

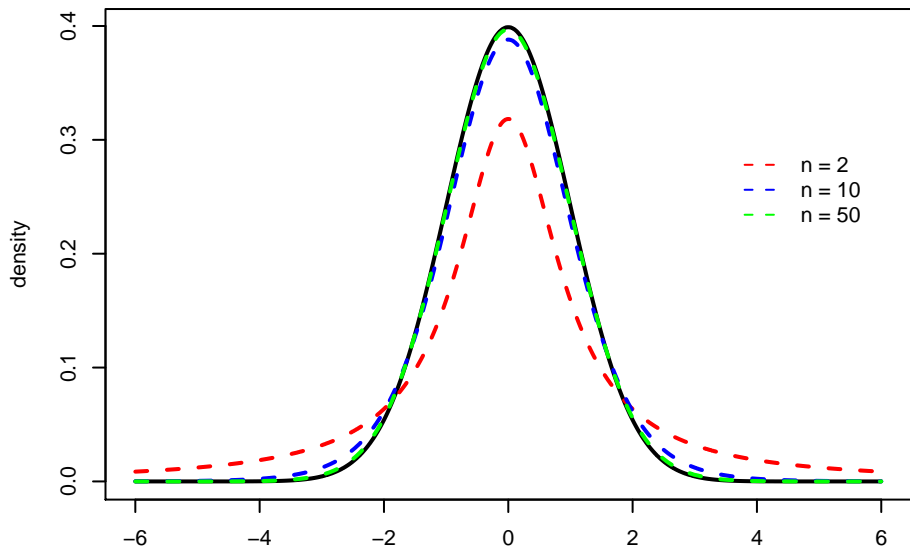
- If $X_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ in a *finite* sample
- Moreover, in a *finite* sample,

$$t\text{-statistic} = \frac{\hat{\mu} - \mu}{\hat{\sigma}/\sqrt{n}} \stackrel{\text{exactly}}{\sim} t_{n-1}$$

where t_{n-1} is the t distribution with $n - 1$ degrees of freedom

- Use t_{n-1} (rather than $\mathcal{N}(0, 1)$) to obtain the critical value for exact confidence intervals
- As n increases, t_{n-1} approaches to $\mathcal{N}(0, 1)$
- Fat tail: more conservative inference with longer CI
- Sum of independent random variables: Bernoulli (Binomial), Exponential (Gamma), Poisson (Poisson), χ^2 (χ^2), etc.

Student's t Distribution



What is Causal Inference?

- Comparison between factual and **counterfactual**
- Incumbency effect:
What would have been the election outcome if a candidate were not an incumbent?
- Resource curse thesis:
What would have been the GDP growth rate without oil?
- Democratic peace theory:
Would the two countries have escalated crisis in the same situation if they were both autocratic?
- SUPPLEMENTARY READING: Holland, P. (1986). Statistics and causal inference. (with discussions) *Journal of the American Statistical Association*, Vol. 81: 945–960.

Defining Causal Effects

- Units: $i = 1, \dots, n$
- “Treatment”: $T_i = 1$ if treated, $T_i = 0$ otherwise
- Observed outcome: Y_i
- Pre-treatment covariates: X_i
- **Potential outcomes**: $Y_i(1)$ and $Y_i(0)$ where $Y_i = Y_i(T_i)$

Voters	Contact	Turnout		Age	Party ID
i	T_i	$Y_i(1)$	$Y_i(0)$	X_i	X_i
1	1	1	?	20	D
2	0	?	0	55	R
3	0	?	1	40	R
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots
n	1	0	?	62	D

- Causal effect: $Y_i(1) - Y_i(0)$

The Key Assumptions

- **No simultaneity** (different from endogeneity)
- **No interference** between units: $Y_i(T_1, T_2, \dots, T_n) = Y_i(T_i)$
- Potential violations:
 - ① spill-over effects
 - ② carry-over effects
- Cluster randomized experiments as a solution (more later)
- Stable Unit Treatment Value Assumption (SUTVA):
no interference + “the same version” of the treatment
- Potential outcome is thought to be fixed: data cannot distinguish fixed and random potential outcomes
- But, potential outcomes across units have a distribution
- Observed outcome is random because the treatment is random
- Multi-valued treatment: more potential outcomes for each unit

Causal Effects of Immutable Characteristics

- “No causation without manipulation” (Holland, 1986)
- Immutable characteristics; gender, race, age, etc.
- What does the causal effect of gender mean?

- Causal effect of having a female politician on policy outcomes (Chattopadhyay and Duflo, 2004 *QJE*)
- Causal effect of having a discussion leader with certain preferences on deliberation outcomes (Humphreys *et al.* 2006 *WP*)
- Causal effect of a job applicant’s gender/race on call-back rates (Bertrand and Mullainathan, 2004 *AER*)

Average Treatment Effects

- Sample Average Treatment Effect (SATE):

$$\frac{1}{n} \sum_{i=1}^n Y_i(1) - Y_i(0)$$

- Population Average Treatment Effect (PATE):

$$\mathbb{E}(Y_i(1) - Y_i(0))$$

- Population Average Treatment Effect for the Treated (PATT):

$$\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$$

- **Causal heterogeneity**: Zero ATE doesn't mean zero effect for everyone!
- Other quantities: Conditional ATE, Quantile Treatment Effects, etc.

Design Considerations

- Randomized experiments
 - Laboratory experiments
 - Survey experiments
 - Field experiments
- Observational studies
- Tradeoff between **internal and external validity**
 - Endogeneity: selection bias
 - Generalizability: sample selection, Hawthorne effects, realism
- “Designing” observational studies
 - Natural experiments (haphazard treatment assignment)
 - Examples: birthdays, weather, close elections, arbitrary administrative rules
- Generalizing experimental results: possible extrapolation
- Bottom line: No study is perfect, statistics is always needed

Truncation by Death

- Units: patients (students)
- Treatment: a new medicine (a new teaching program)
- Outcome: blood pressure (the end-of-year test score)
- Truncation: some patients die (some students drop out)
- Blood pressure (test score) **undefined** for the dead (drop-outs)!

- Randomized evaluation of Mexican health care
- Units: individuals
- Treatment: universal health insurance
- Outcome: satisfaction with the received health care
- Truncation: some have not been to clinics or hospitals

- SUPPLEMENTAL READING: Rubin, D. B. (2006). Causal Inference Through Potential Outcomes and Principal Stratification: Application to Studies with “Censoring” Due to Death *Statistical Science*, Vol. 21: 299 – 309. Skip Sections 4 – 6.

Principal Stratification

- Comparison between the dead and the alive is invalid
 - Potential selection bias
 - Causal effects need to be defined at the unit level

- Binary treatment: $T_i \in \{0, 1\}$
- Potential truncation variable: $W_i(1), W_i(0)$
- Observed truncation variable: $W_i = W_i(T_i)$
- Potential outcomes: $Y_i(0, 0), Y_i(1, 0)$
- $Y_i(0, 1), Y_i(1, 1)$ do not exist
- Observed outcome: $Y_i = Y_i(T_i, W_i)$ for $W_i = 0$

- Causal effect: $Y_i(1, 0) - Y_i(0, 0)$ for “always-survivors”
- Not defined for other types (**principal strata**)

- LECTURE NOTES: “Statistical Framework of Causal Inference”

(Classical) Randomized Experiments

- Units: $i = 1, \dots, n$
- May constitute a simple random sample from a population
- Treatment: $T_i \in \{0, 1\}$
- Outcome: $Y_i = Y_i(T_i)$
- Complete randomization of the treatment assignment
- Exactly n_1 units receive the treatment
- $n_0 = n - n_1$ units are assigned to the control group
- **Assumption:** for all $i = 1, \dots, n$, $\sum_{i=1}^n T_i = n_1$ and

$$(Y_i(1), Y_i(0)) \perp\!\!\!\perp T_i, \quad \Pr(T_i = 1) = \frac{n_1}{n}$$

- Estimand = SATE or PATE
- Estimator = Difference-in-means:

$$\hat{\tau} \equiv \frac{1}{n_1} \sum_{i=1}^n T_i Y_i - \frac{1}{n_0} \sum_{i=1}^n (1 - T_i) Y_i$$

Estimation of Average Treatment Effects

- Key idea (Neyman 1923): Randomness comes from treatment assignment (plus sampling for PATE) alone
- Design-based (randomization-based) rather than model-based
- Statistical properties of $\hat{\tau}$ based on design features
- Define $\mathcal{O} \equiv \{Y_i(0), Y_i(1)\}_{i=1}^n$
- Unbiasedness (over repeated treatment assignments):

$$\begin{aligned}\mathbb{E}(\hat{\tau} \mid \mathcal{O}) &= \frac{1}{n_1} \sum_{i=1}^n \mathbb{E}(T_i \mid \mathcal{O}) Y_i(1) - \frac{1}{n_0} \sum_{i=1}^n \{1 - \mathbb{E}(T_i \mid \mathcal{O})\} Y_i(0) \\ &= \frac{1}{n} \sum_{i=1}^n (Y_i(1) - Y_i(0)) \\ &= \text{SATE}\end{aligned}$$

Variance for Within-Sample Inference

- Exact variance of $\hat{\tau}$:

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) = \frac{1}{n} \left(\frac{n_0}{n_1} S_1^2 + \frac{n_1}{n_0} S_0^2 + 2S_{01} \right),$$

where for $t = 0, 1$,

$$S_t^2 = \frac{1}{n-1} \sum_{i=1}^n (Y_i(t) - \overline{Y(t)})^2 \quad \text{sample variance of } Y_i(t)$$

$$S_{01} = \frac{1}{n-1} \sum_{i=1}^n (Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)}) \quad \text{sample covariance}$$

- The variance is NOT *identifiable*

Randomization Inference with the Difference-in-Means

- 1 Let $X_i = Y_i(1) + n_1 Y_i(0)/n_0$ and $D_i = nT_i/n_1 - 1$, and write

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) = \frac{1}{n^2} \mathbb{E} \left[\left(\sum_{i=1}^n D_i X_i \right)^2 \mid \mathcal{O} \right]$$

- 2 Show

$$\begin{aligned} \mathbb{E}(D_i \mid \mathcal{O}) &= 0, & \mathbb{E}(D_i^2 \mid \mathcal{O}) &= \frac{n_0}{n_1}, \\ \mathbb{E}(D_i D_j \mid \mathcal{O}) &= -\frac{n_0}{n_1(n-1)} \end{aligned}$$

- 3 Use 1 and 2 to show,

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) = \frac{n_0}{n(n-1)n_1} \sum_{i=1}^n (X_i - \bar{X})^2$$

- 4 Substitute the potential outcome expressions for X_i

Some Remarks on Finite Sample Inference

- The usual variance estimator is conservative on average:

$$\mathbb{V}(\hat{\tau} \mid \mathcal{O}) \leq \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

- Under the constant additive unit causal effect assumption, i.e., $Y_i(1) - Y_i(0) = c$ for all i ,

$$S_{01} = \frac{1}{2}(S_1^2 + S_0^2)$$

- The optimal treatment assignment:

$$n_1^{opt} = \frac{n}{1 + S_0/S_1}, \quad n_0^{opt} = \frac{n}{1 + S_1/S_0}$$

Population Inference

- Now assume that units are randomly sampled from a population
- Unbiasedness (over repeated sampling):

$$\begin{aligned}\mathbb{E}\{\mathbb{E}(\hat{\tau} \mid \mathcal{O})\} &= \mathbb{E}(\text{SATE}) \\ &= \mathbb{E}(Y_i(1) - Y_i(0)) \\ &= \text{PATE}\end{aligned}$$

- Exact variance

$$\begin{aligned}\mathbb{V}(\hat{\tau}) &= \mathbb{V}(\mathbb{E}(\hat{\tau} \mid \mathcal{O})) + \mathbb{E}(\mathbb{V}(\hat{\tau} \mid \mathcal{O})) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\end{aligned}$$

where σ_t^2 is the population variance of $Y_i(t)$ for $t = 0, 1$

Asymptotic Population Inference

- Hold $k = n_1/n$ constant
- Rewrite the difference-in-means estimator as

$$\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \underbrace{\left(\frac{T_i Y_i(1)}{k} - \frac{(1 - T_i) Y_i(0)}{1 - k} \right)}_{i.i.d. \text{ with mean PATE \& variance } n\mathbb{V}(\hat{\tau})}$$

- Consistency:

$$\hat{\tau} \xrightarrow{p} \text{PATE}$$

- Asymptotic normality:

$$\sqrt{n}(\hat{\tau} - \text{PATE}) \xrightarrow{d} \mathcal{N}\left(0, \frac{\sigma_1^2}{k} + \frac{\sigma_0^2}{1 - k}\right)$$

- $(1 - \alpha) \times 100\%$ Confidence intervals:

$$[\hat{\tau} - \text{s.e.} \times Z_{\alpha/2}, \hat{\tau} + \text{s.e.} \times Z_{\alpha/2}]$$

A Model-based View

- A random sample of n_1 units from the “treatment” population of infinite size
- A random sample of n_0 units from the “control” population of infinite size
- The randomization of the treatment implies that two populations are identical except the receipt of the treatment
- The difference in the population means = PATE
- The lack of randomization means no causal interpretation but the difference in the population means can always be estimated
- Unbiased estimator from the model-based sample surveys:

$$\hat{\tau} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i} - \frac{1}{n_0} \sum_{i=1}^{n_0} Y_{0i}$$

- Variance is the same: $\mathbb{V}(\hat{\tau}) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}$

Paul the Octopus and Statistical Hypothesis Tests



- 2010 World Cup
 - Group: **Germany** vs Australia
 - Group: Germany vs **Serbia**
 - Group: Ghana vs **Germany**
 - Round of 16: **Germany** vs England
 - Quarter-final: Argentina vs **Germany**
 - Semi-final: Germany vs **Spain**
 - 3rd place: Uruguay vs **Germany**
 - Final: Netherlands vs **Spain**

- Question: Did Paul the Octopus get lucky?
- Suppose that Paul is randomly choosing winner
- Then, # of correct answers \sim Binomial(8, 0.5)
- The probability that Paul gets them all correct: $\frac{1}{2^8} \approx 0.004$
- Tie is possible in group rounds: $\frac{1}{3^3} \times \frac{1}{2^5} \approx 0.001$
- Conclusion: Paul may be a prophet

What are Statistical Hypothesis Tests?

- Probabilistic “Proof by contradiction”
- ① Choose a **null hypothesis** (H_0) and an **alternative hypothesis** (H_1)
- ② Choose a **test statistic** Z
- ③ Derive the sampling distribution of Z under H_0 (**reference distribution**)
- ④ Is the observed value of Z likely to occur under H_0 ?
 - Yes – Retain H_0 (\neq accept H_0)
 - No – Reject H_0

More Data about Paul

- UEFA Euro 2008

- Group: **Germany** vs Poland
- **Group**: Croatia vs **Germany**
- Group: Austria vs **Germany**
- Quarter-final: Portugal vs **Germany**
- Semi-final: **Germany** vs Turkey
- **Final**: **Germany** vs Spain

- A total of 14 matches

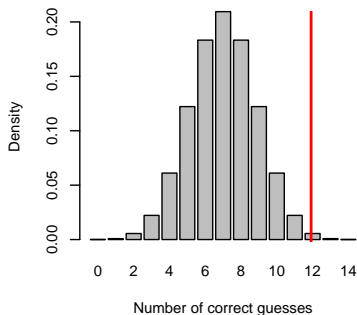
- 12 correct guesses

- **p-value**: Probability that under the null you observe something at least as extreme as what you actually observed

- $\Pr(\{12, 13, 14\}) \approx 0.001$

- R: `pbinom(12, size = 14, prob = 0.5, lower.tail = FALSE)`

Reference distribution: Binom(14, 0.5)



Paul's Rival, Mani the Parakeet



● 2010 World Cup

- Quarter-final: Netherlands vs Brazil
- Quarter-final: Uruguay vs Ghana
- Quarter-final: Argentina vs Germany
- Quarter-final: Paraguay vs Spain
- Semi-final: Uruguay vs Netherlands
- Semi-final: Germany vs Spain
- Final: Netherlands vs Spain

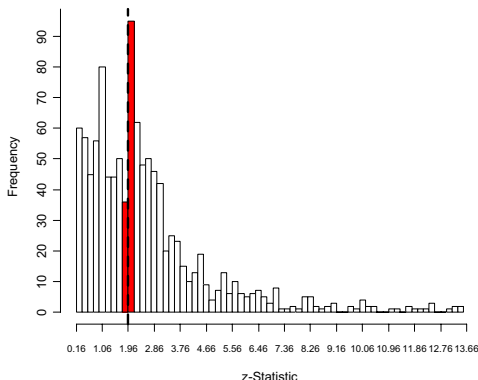
- Mani did pretty good too: p -value is 0.0625
- Danger of multiple testing \implies false discovery
- Take 10 animals with no forecasting ability. What is the chance of getting p -value less than 0.05 at least once?

$$1 - 0.95^{10} \approx 0.4$$

- If you do this with enough animals, you will find another Paul

False Discovery and Publication Bias

- Publication bias (Gerber and Malhotra, *QJPS* 2008):



- Statistical methods to control for **false discovery rate** (FDR):

$$\mathbb{E} \left\{ \frac{\# \text{ of false rejections}}{\max(\text{total } \# \text{ of rejections}, 1)} \right\}$$

Error and Power of Hypothesis Test

- Two types of errors:

	Reject H_0	Retain H_0
H_0 is true	Type I error	Correct
H_0 is false	Correct	Type II error

- Hypothesis tests control the probability of Type I error
- They do not control the probability of Type II error
- Tradeoff between the two types of error
- **Size** of test: probability that the null is rejected when it is true
- **Power** of test: probability that a test rejects the null
- Typically, we want a most powerful test with the proper size

p -value and Statistical Significance

- p -value is the probability, computed under H_0 , of observing a value of the test statistic at least as extreme as its observed value
- A smaller p -value presents stronger evidence against H_0
- p -value less than α indicates **statistical significance** \leftrightarrow α -level test
- α is called the significance level

- p -value is NOT the probability that H_0 (H_1) is true (false)
- A large p -value can occur either because H_0 is true or because H_0 is false but the test is not powerful
- The statistical significance indicated by the p -value does not necessarily imply scientific significance

- **Inverting the hypothesis test** to obtain confidence intervals
- Typically better to present confidence intervals than p -values

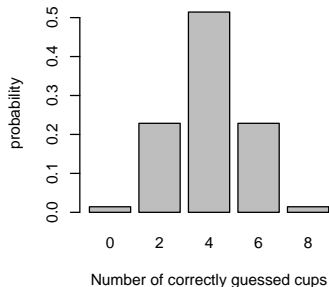
Lady Tasting Tea

- Does tea taste different depending on whether the tea was poured into the milk or whether the milk was poured into the tea?
- 8 cups; $n = 8$
- Randomly choose 4 cups into which pour the tea first ($T_i = 1$)
- Null hypothesis: the lady cannot tell the difference
- Sharp null – $H_0 : Y_i(1) = Y_i(0)$ for all $i = 1, \dots, 8$
- Statistic: the number of correctly classified cups
- The lady classified all 8 cups correctly!
- Did this happen by chance?

- SUPPLEMENTARY READING: Ho and Imai (2006). “Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election.” *J. of the Amer. Stat. Assoc.*

Randomization Test (Fisher's Exact Test)

cups	guess	actual	scenarios	...
1	M	M	T	T
2	T	T	T	T
3	T	T	T	T
4	M	M	T	M
5	M	M	M	M
6	T	T	M	M
7	T	T	M	T
8	M	M	M	M
correctly guessed		8	4	6



- ${}_8C_4 = 70$ ways to do this and each arrangement is equally likely
- What is the p -value?
- No assumption, but the sharp null may be of little interest

One-Sample Test

- Looks and politics: *Todorov et al. Science*



Which person is the more competent?

- $H_0: p_{comp} = 0.5$ and $H_1: p_{comp} \neq 0.5$
- Test statistic: sample mean \hat{p}_{comp}
- Exact reference distribution: $\text{Binom}(n, 0.5)$
- Asymptotic reference distribution:

$$\frac{\hat{p}_{comp} - 0.5}{0.5/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

Two-Sample Test

- $H_0 : \text{PATE} = \tau_0$ and $H_1 : \text{PATE} \neq \tau_0$
- Asymptotic reference distribution:

$$Z\text{-statistic} = \frac{\hat{\tau} - \tau_0}{\text{s.e.}} \xrightarrow{D} \mathcal{N}(0, 1)$$

- Is Z_{obs} unusual under the null?
 - Reject the null when $|Z_{obs}| > z_{\alpha/2}$
 - Retain the null when $|Z_{obs}| \leq z_{\alpha/2}$
- If we assume $Y_i(1) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_1, \sigma_1^2)$ and $Y_i(0) \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_0, \sigma_0^2)$, then

$$t\text{-statistic} = \frac{\hat{\tau} - \tau_0}{\text{s.e.}} \sim t_\nu$$

where ν is given by a complex formula (Behrens-Fisher problem)

Power Analysis

- Null hypotheses are often uninteresting
- But, hypothesis testing may indicate the strength of evidence for or against your theory
- Power analysis: What sample size do I need in order to detect a certain departure from the null?
- Power = $1 - \Pr(\text{Type II error})$
- Three steps (for sample mean)
 - 1 Suppose $\mu = \mu^*$ which implies $\bar{X} \sim \mathcal{N}(\mu^*, \mathbb{V}(X)/n)$
 - 2 Calculate the rejection probability noting that we reject $H_0 : \mu = \mu_0$ if $|\bar{X}| > \mu_0 + z_{\alpha/2} \times \text{s.e.}$
 - 3 Find the smallest n such that this rejection probability equals a prespecified level

One-Sided Example

- $H_0 : p = p_0$ and $H_a : p > p_0$
- $\bar{X} \sim \mathcal{N}(p^*, p^*(1 - p^*)/n)$
- Reject H_0 if $\bar{X} > p_0 + z_{\alpha/2} \times \sqrt{p_0(1 - p_0)/n}$

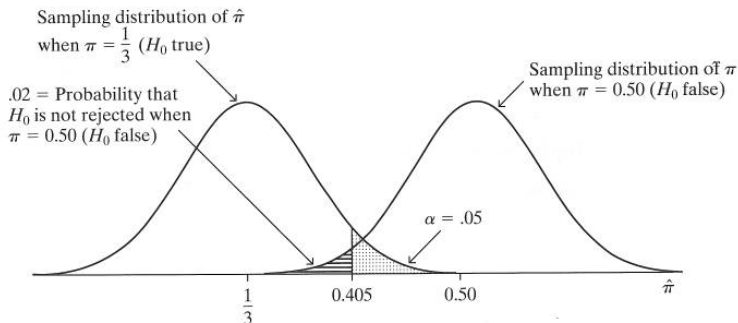
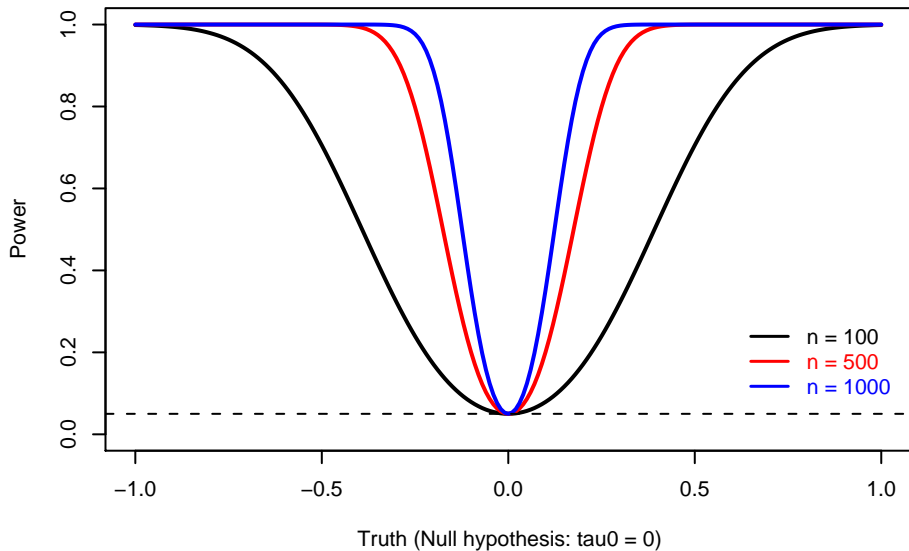


FIGURE 6.11: Calculation of $P(\text{Type II Error})$ for Testing $H_0: \pi = 1/3$ against $H_a: \pi > 1/3$ at $\alpha = 0.05$ Level, when True Proportion is $\pi = 0.50$. A Type II error occurs if $\hat{\pi} < 0.405$, since then $P\text{-value} > 0.05$ even though H_0 is false.

Power Function ($\sigma_0^2 = \sigma_1^2 = 1$ and $n_1 = n_0$)



Identification vs. Estimation

- Identification: How much can you learn about the estimand if you had an infinite amount of data?
- Estimation: How much can you learn about the estimand from a finite sample?
- Identification precedes estimation
- *Partial* (sharp bounds) vs. *Point* identification (point estimates)
 - ① What can be learned without any assumption other than the ones which we know are satisfied by the research design?
 - ② What is a minimum set of assumptions required for point identification?
 - ③ Can we characterize identification region if we relax some or all of these assumptions?
- **Law of Decreasing Credibility** (Manski): The credibility of inference decreases with the strength of the assumptions maintained

Identification of the Average Treatment Effect

- Assumption 1: Overlap (i.e., no extrapolation)

$$0 < \Pr(T_i = 1 \mid X_i = x) < 1 \text{ for any } x \in \mathcal{X}$$

- Assumption 2: Ignorability (exogeneity, unconfoundedness, no omitted variable, selection on observables, etc.)

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid X_i = x \text{ for any } x \in \mathcal{X}$$

- Under these assumptions, we have **nonparametric identification**:

$$\tau = \mathbb{E}\{\mu(1, X_i) - \mu(0, X_i)\}$$

where $\mu(t, x) = \mathbb{E}(Y_i \mid T_i = t, X_i = x)$

Sharp Bounds

- In randomized experiments, these assumptions are satisfied
- In observational studies, their validity is in doubt
- What happens if the assumptions are violated?
- Identification analysis without assuming a model

- A special case with binary outcome:

$$\begin{aligned} &[-\Pr(Y_i = 0 \mid T_i = 1, X_i = x)e(x) - \Pr(Y_i = 1 \mid T_i = 0, X_i = x)\{1 - e(x)\}, \\ &\Pr(Y_i = 1 \mid T_i = 1, X_i = x)e(x) + \Pr(Y_i = 0 \mid T_i = 0, X_i = x)\{1 - e(x)\}] \end{aligned}$$

where $e(x) = \Pr(T_i = 1 \mid X_i = x)$ is called **propensity score**

- The width of the bounds is 1: “A glass is half empty/full”

Characterization of Asymptotic Bias

- Estimation of PATT: $\mathbb{E}(Y_i(1) - Y_i(0) \mid T_i = 1)$
- Decomposition of asymptotic bias:

$$\begin{aligned} & \mathbb{E}(Y_i(0) \mid T_i = 1) - \mathbb{E}\{\mu(0, X_i)\} \\ = & \int_{S_1 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 1, X_i = x) dF_{X_i \mid T_i=1}(x) \\ & - \int_{S_0 \setminus S} \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x) dF_{X_i \mid T_i=0}(x) \\ + & \int_S \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x) d\{F_{X_i \mid T_i=1}(x) - F_{X_i \mid T_i=0}(x)\} \\ + & \int_S \{\mathbb{E}(Y_i(0) \mid T_i = 1, X_i = x) - \mathbb{E}(Y_i(0) \mid T_i = 0, X_i = x)\} dF_{X_i \mid T_i=1}(x) \end{aligned}$$

- Statistical methods under exogeneity: matching and regression
- Research designs to overcome exogeneity: “natural” experiments

Concluding Remarks

1 Estimation and Inference

- Plug-in principle: use sample analogue to estimate a population parameter
- Design-based vs. model-based inference: where does the randomness come from?
- Importance of uncertainty estimation: confidence intervals

2 Hypothesis Testing

- Stochastic proof by contradiction
- Null hypothesis: Do not accept it. Is it a quantity of interest?
- Power analysis: important at the planning and analysis stages
- Danger of multiple testing and data snooping

3 Identification

- Identification precedes estimation: importance of research design
- Partial identification with weaker assumptions
- Identification of causal effects vs. descriptive parameters