# Basic Principles of Statistical Inference

Kosuke Imai

Department of Politics
Princeton University

POL572 Quantitative Analysis II
Spring 2016

# 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 collect 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:

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

- Statistical methods are no substitute for good research design

# Sample Surveys

# 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 $i = 1, \ldots, N$
- Sampling (binary) indicator: $Z_1, \ldots, Z_N$
- Assumption: $\sum_{i=1}^N Z_i = n$ and $\Pr(Z_i = 1) = n/N$ for all $i$
- # of combinations: $\binom{N}{n} = \frac{N!}{n!(N-n)!}$
- Estimand = population mean vs. Estimator = sample mean:

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

# Estimation of Population Mean

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

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

  where $S^2 = \sum_{i=1}^{N}(X_i - \overline{X})^2/(N-1)$ 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

# Some VERY Important Identities in Statistics

1. $\mathbb{V}(X) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$
2. $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$
3. Law of Iterated Expectation:

$$\mathbb{E}(X) \;=\; \mathbb{E}\{\mathbb{E}(X \mid Y)\}$$

4. Law of Total Variance:

$$\mathbb{V}(X) \;=\; \underbrace{\mathbb{E}\{\mathbb{V}(X \mid Y)\}}_{\text{within}-\text{group variance}} + \underbrace{\mathbb{V}\{\mathbb{E}(X \mid Y)\}}_{\text{between}-\text{group variance}}$$

5. Mean Squared Error Decomposition:

$$\mathbb{E}\{(\hat{\theta} - \theta)^2\} \;=\; \underbrace{\{\mathbb{E}(\hat{\theta} - \theta)\}^2}_{\text{bias}^2} + \underbrace{\mathbb{V}(\hat{\theta})}_{\text{variance}}$$

# Analytical Details of Randomization Inference

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

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

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

4. Finally, we proceed as follows:

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

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

# Inverse Probability Weighting

- Unequal sampling probability: $\Pr(Z_i = 1) = \pi_i$ for each $i$
- We still randomly sample $n$ units from the population of size $N$ where $\sum_{i=1}^N Z_i = n$ implying $\sum_{i=1}^N \pi_i = n$
- Oversampling of minorities, difficult-to-reach individuals, etc.
- Sampling weights $=$ inverse of sampling probability
- Horvitz-Thompson estimator:

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

- Unbiasedness: $\mathbb{E}(\tilde{x}) = \overline{X}$
- Design-based variance is complicated but available
- Háyek estimator (biased but possibly more efficient):

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

- Unknow sampling probability $\leadsto$ post-stratification

# Model-Based Inference

- An infinite population characterized by a probability *model*
  - Nonparametric $\mathcal{F}$
  - Parametric $\mathcal{F}_\theta$ (e.g., $\mathcal{N}(\mu, \sigma^2)$)
- A simple random sample of size $n$: $X_1, \ldots, X_n$
- 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:

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

where $\sigma^2 = \mathbb{V}(X_i)$

# (Weak) Law of Large Numbers (LLN)

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

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

  where "$\xrightarrow{p}$" denotes the convergence in probability, i.e., if $X_n \xrightarrow{p} x$, then

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

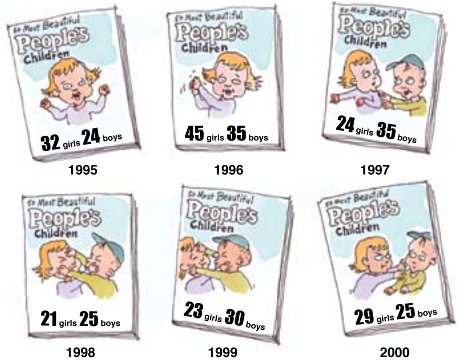
- If $X_n \xrightarrow{p} x$, then for any continuous function $f(\cdot)$, we have

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

- 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)




- 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 dupilicates: 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 $\mu$ and finite variance $\sigma^2$, then

$$\underbrace{\frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}}}_{z\text{-score of sample mean}} \quad \xrightarrow{d} \mathcal{N}(0, 1)$$

  where "$\xrightarrow{d}$" represents the convergence in distribution, i.e., if $X_n \xrightarrow{d} X$, then

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

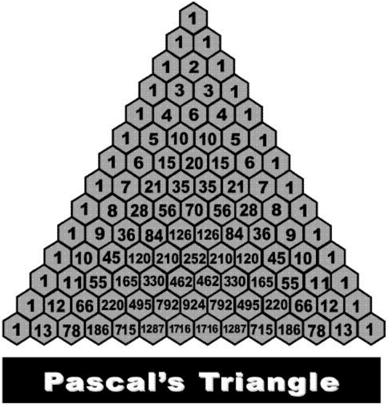
  with $P(X \leq x)$ being continuous at every $x$

- 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

# CLT in Action




- $n^{th}$ 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, Boston Museum of Science, or just check out YouTube

# Asymptotic Properties of the Sample Mean

- The Model: $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{F}_{\mu,\sigma^2}$
- LLN implies consistency:

$$\hat{\mu} \;=\; \overline{X}_n \;\xrightarrow{p}\; \mu$$

- CLT implies asymptotic normality:

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

  But, $\sigma$ 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 $\sigma^2$ (LLN)

# Asymptotic Confidence Intervals

- Putting together, we have:

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

- 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:

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

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

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

$$\mathrm{CI}_{1-\alpha} \;=\; [\hat{\mu} - z_{\alpha/2} \times \mathrm{s.e.}, \quad \hat{\mu} + z_{\alpha/2} \times \mathrm{s.e.}]$$

  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: $\Pr(\mu \in \mathrm{CI}_{1-\alpha}) \xrightarrow{p} 1-\alpha$
- Asymptotic inference $=$ approximate inference

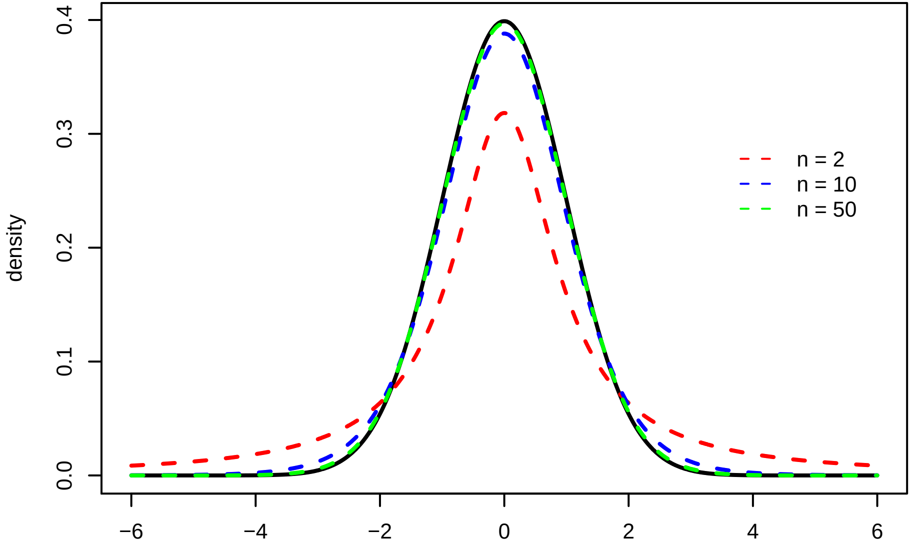
# Exact Inference with Normally Distributed Data

- Sometimes, exact model-based inference is possible
- If $X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$, then $\hat{\mu} \sim \mathcal{N}(\mu, \sigma^2/n)$ in a *finite* sample
- Moreover, in a *finite* sample,

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

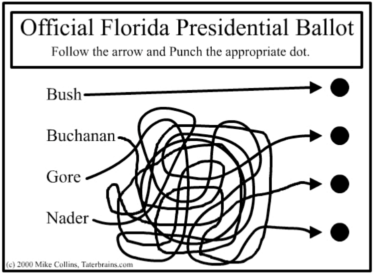
  where $t_{n-1}$ is the t distribution with $n-1$ degrees of freedom
- Use $t_{n-1}$ (rather than $\mathcal{N}(0,1)$) to obtain the critical value for exact confidence intervals
- As $n$ increases, $t_{n-1}$ approaches to $\mathcal{N}(0,1)$
- Fat tail: more conservative inference with wider CI
- Sum of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

# Student's $t$ Distribution



# Application: Presidential Election Polling

- 2000 Butterfly ballot debacle: Oops, we have this system called electoral college!




- National polls $\Longrightarrow$ state polls
- Forecasting fun: political methodologists, other "statisticians"
- Idea: estimate probability that each state is won by a candidate and then aggregate electoral votes
- Quantity of interest: Probability of a candidate winning the election

# Simple Model-Based Inference

- Setup: $n_{jk}$ respondents of poll $j$ from state $k$
- Model for # of Obama supporters in poll $j$ and state $k$:

$$X_{jk} \overset{\text{indep.}}{\sim} \text{Binom}(n_{jk}, p_k)$$

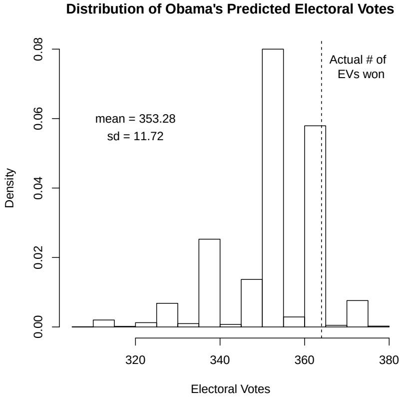
- Parameters of interest: $\theta = \{p_1, p_2, \ldots, p_{51}\}$
- Popular methods of inference:
  1. Method of moments (MM) $\rightarrow$ solve the moment equation sample moments$(X)$ $=$ population moments$(\theta)$
  2. Maximum likelihood (ML) $\rightarrow$ maximize the likelihood $f(X \mid \theta)$
  3. Bayesian inference $\rightarrow$ derive the posterior of parameters

$$f(\theta \mid X) = \frac{\overbrace{f(X \mid \theta)}^{\text{likelihood}} \times \overbrace{f(\theta)}^{\text{prior}}}{\underbrace{f(X)}_{\text{marginal likelihood} = \int f(X|\theta)f(\theta)d\theta}} \propto f(X \mid \theta)\, f(\theta)$$

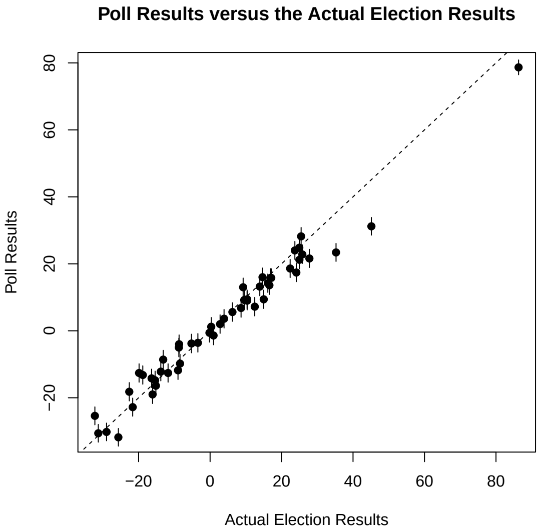
- In this case, MM and ML give $\hat{p}_k = \sum_{j=1}^{J_k} X_{jk} / \sum_{j=1}^{J_k} n_{jk}$

# Estimated Probability of Obama Victory in 2008

- Estimate $p_k$ for each state
- Simulate $M$ elections using $\hat{p}_k$ and its standard error:
  1. for state $k$, sample Obama's voteshare from $\mathcal{N}(\hat{p}_k, \widehat{\mathbb{V}(\hat{p}_k)})$
  2. collect all electoral votes from winning states
- Plot $M$ draws of total electoral votes



# Nominal vs. Actual Coverage




- Coverage: 55%
- Bias: 1 ppt.
- Bias-adjusted coverage: 60%
- Still significant undercoverage

# Key Points

- Random sampling enables statistical inference
- Design-based vs. Model-based inference
  1. Design-based: random sampling as basis for inference
  2. Model-based: probability model as basis for inference
- Sampling weights: inverse probability weighting
- Challenges of survey research:
  - cluster sampling, multi-stage sampling $\Longrightarrow$ loss of efficiency
  - stratified sampling
  - unit non-response
  - non-probability sampling $\Longrightarrow$ model-based inference
  - item non-response, social desirability bias, etc.

# Causal Inference

# What is Causal Inference?

- Comparison between factual and counterfactual for each unit
- 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: $i = 1, \ldots, n$
- "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

- The notation implies three assumptions:
  1. No simultaneity (different from endogeneity)
  2. No interference between units: $Y_i(T_1, T_2, \ldots, T_n) = Y_i(T_i)$
  3. Same version of the treatment
- Stable Unit Treatment Value Assumption (SUTVA)
- Potential violations:
  1. feedback effects
  2. spill-over effects, carry-over effects
  3. different treatment administration

- Potential outcome is thought to be "fixed": data cannot distinguish fixed and random potential outcomes
- 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*)

- Problem: confounding

# 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):

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

- Treatment effect heterogeneity: Zero ATE doesn't mean zero effect for everyone! $\Longrightarrow$ Conditional ATE
- Other quantities: 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

# (Classical) Randomized Experiments

- Units: $i = 1, \ldots, n$
- 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 $i = 1, \ldots, n$, $\sum_{i=1}^{n} T_i = n_1$ 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:

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

# Unbiased 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):

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

# Randomization Inference for SATE

- 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 $t = 0, 1$,

$$\begin{aligned} S_t^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(t) - \overline{Y(t)})^2 \quad \text{sample variance of } Y_i(t) \\ S_{01} &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i(0) - \overline{Y(0)})(Y_i(1) - \overline{Y(1)}) \quad \text{sample covariance} \end{aligned}$$

- The variance is NOT *identifiable*

- 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 $i$,

$$S_{01} = \frac{1}{2}(S_1^2 + S_0^2) \quad \text{and} \quad \mathbb{V}(\hat{\tau} \mid \mathcal{O}) = \frac{S_1^2}{n_1} + \frac{S_0^2}{n_0}$$

- The optimal treatment assignment rule:

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

# Details of Variance Derivation

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

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

2. Show

$$\mathbb{E}(D_i \mid \mathcal{O}) = 0, \quad \mathbb{E}(D_i^2 \mid \mathcal{O}) = \frac{n_0}{n_1},$$

$$\mathbb{E}(D_i D_j \mid \mathcal{O}) = -\frac{n_0}{n_1(n-1)}$$

3. Use ❶ and ❷ to show,

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

4. Substitute the potential outcome expressions for $X_i$

# Randomization Inference for PATE

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

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

- 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 $\sigma_t^2$ is the population variance of $Y_i(t)$ for $t = 0, 1$

# Asymptotic Inference for PATE

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

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

- 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:

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

# Model-based Inference about PATE

- 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

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

# Identification vs. Estimation

- Observational studies $\Longrightarrow$ No randomization of treatment
- Difference in means between two populations can still be estimated without bias
- Valid inference for ATE requires additional assumptions
- Law of Decreasing Credibility (Manski): The credibility of inference decreases with the strength of the assumptions maintained

- 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

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

-

# Partial Identification

- 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?

- ATE with binary outcome:

$$[-\Pr(Y_i = 0 \mid T_i = 1, X_i = x)\pi(x) - \Pr(Y_i = 1 \mid T_i = 0, X_i = x)\{1 - \pi(x)\},$$
$$\Pr(Y_i = 1 \mid T_i = 1, X_i = x)\pi(x) + \Pr(Y_i = 0 \mid T_i = 0, X_i = x)\{1 - \pi(x)\}]$$

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

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

# Application: List Experiment

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the control group

```
Now I'm going to read you three things that sometimes
make people angry or upset.  After I read all three,
just tell me HOW MANY of them upset you.  (I don't
want to know which ones, just how many.)

(1) the federal government increasing the tax on
   gasoline;
(2) professional athletes getting million-dollar-plus
   salaries;
(3) large corporations polluting the environment.
```

# Application: List Experiment

- The 1991 National Race and Politics Survey (Sniderman et al.)
- Randomize the sample into the treatment and control groups
- The script for the treatment group

```
Now I'm going to read you four things that sometimes
make people angry or upset.  After I read all four,
just tell me HOW MANY of them upset you.  (I don't
want to know which ones, just how many.)

(1) the federal government increasing the tax on
   gasoline;
(2) professional athletes getting million-dollar-plus
   salaries;
(3) large corporations polluting the environment;
(4) a black family moving next door to you.
```

# Identification Assumptions and Potential Outcomes

- Identification assumptions:
  1. No Design Effect: The inclusion of the sensitive item does not affect answers to control items
  2. No Liars: Answers about the sensitive item are truthful
- Define a type of each respondent by
  - total number of yes for control items $Y_i(0)$
  - truthful answer to the sensitive item $Z_i^*$
- Under the above assumptions, $Y_i(1) = Y_i(0) + Z_i^*$
- A total of $(2 \times (J + 1))$ types

# Example with 3 Control Items

- *Joint distribution* of $\pi_{yz} = (Y_i(0) = y, Z_i^* = z)$ is identified:

| $Y_i$ | Treatment group | Control group |
|---|---|---|
| 4 | (3,1) | |
| 3 | (2,1) (3,0) | (3,1) (3,0) |
| 2 | (1,1) (2,0) | (2,1) (2,0) |
| 1 | ~~(0,1)~~ ~~(1,0)~~ | (1,1) ~~(1,0)~~ |
| 0 | ~~(0,0)~~ | ~~(0,1)~~ ~~(0,0)~~ |

- Testing the validity of the identification assumptions: if the assumptions are valid, $\pi_{yz}$ should be positive for all $y$ and $z$
- Suppose that a negative value of $\hat{\pi}_{yz}$ is observed. Did this happen by chance?
- Statistical hypothesis test (next topic)

# Key Points

- Causal inference is all about predicting counter-factuals
- Association (comparison between treated and control groups) is not causation (comparison between factuals and counterfactuals)

- Randomization of treatment eliminates both observed and unobserved confounders
- Design-based vs. model-based inference

- Observational studies $\Longrightarrow$ identification problem
- Importance of research design: What is your identification strategy?

# Statistical Hypothesis Test

# 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

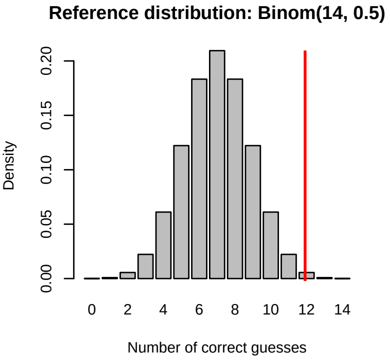
- Question: Did Paul the Octopus get lucky?
- Suppose that Paul is randomly choosing winner
- Then, # of correct answers $\sim \text{Binomial}(8, 0.5)$
- The probability that Paul gets them all correct: $\frac{1}{2^8} \approx 0.004$
- Tie is possible in group rounds: $\frac{1}{3^3} \times \frac{1}{2^5}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

# What are Statistical Hypothesis Tests?

- Probabilistic "Proof by contradiction"
- General procedure:
  1. Choose a null hypothesis ($H_0$) and an alternative hypothesis ($H_1$)
  2. Choose a test statistic $Z$
  3. Derive the sampling distribution (or reference distribution) of $Z$ under $H_0$
  4. Is the observed value of $Z$ likely to occur under $H_0$?
     - Yes $\Longrightarrow$ Retain $H_0$ ($\neq$ accept $H_0$)
     - No $\Longrightarrow$ 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
- $\Pr(\{12, 13, 14\}) \approx 0.001$
- In **R**: `pbinom(12, size = 14, prob = 0.5, lower.tail = FALSE)`

# $p$-value and Statistical Significance

- $p$-value: 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 $\alpha$ indicates statistical significance at the significance level $\alpha$

- $p$-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

# One-Sample Test

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

**Which person is the more competent?**

- $p =$ probability that a more competent politician wins
- $H_0$: $p = 0.5$ and $H_1 : p > 0.5$
- Test statistic $\hat{p} =$ sample proportion
- Exact reference distribution: $\hat{p} \sim \text{Binom}(n, 0.5)$
- Asymptotic reference distribution via CLT:

$$Z\text{-statistic} = \frac{\hat{p} - 0.5}{\text{s.e.}} = \frac{\hat{p} - 0.5}{0.5/\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

# Two-Sample Test

- $H_0 : \text{PATE} = \tau_0$ and $H_1 : \text{PATE} \neq \tau_0$
- Difference-in-means estimator: $\hat{\tau}$
- Asymptotic reference distribution:

$$Z\text{-statistic} = \frac{\hat{\tau} - \tau_0}{\text{s.e.}} = \frac{\hat{\tau} - \tau_0}{\sqrt{\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_0^2}{n_0}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- Is $Z_{obs}$ unusual under the null?
  - Reject the null when $|Z_{obs}| > z_{1-\alpha/2}$
  - Retain the null when $|Z_{obs}| \leq z_{1-\alpha/2}$
- 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 $\nu$ is given by a complex formula (Behrens-Fisher problem)

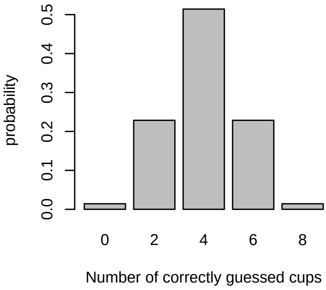
# Lady Tasting Tea

- 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 $i = 1, \ldots, 8$
- Statistic: the number of correctly classified cups
- The lady classified all 8 cups correctly!
- Did this happen by chance?

- Example: 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

# 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 (level) 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

# 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 $= 1 - \Pr(\text{Type II error})$
- Four steps:
  1. Specify the null hypothesis to be tested and the significance level $\alpha$
  2. Choose a true value for the parameter of interest and derive the sampling distribution of test statistic
  3. Calculate the probability of rejecting the null hypothesis under this sampling distribution
  4. Find the smallest sample size such that this rejection probability equals a prespecified level

# One-Sided Test Example

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

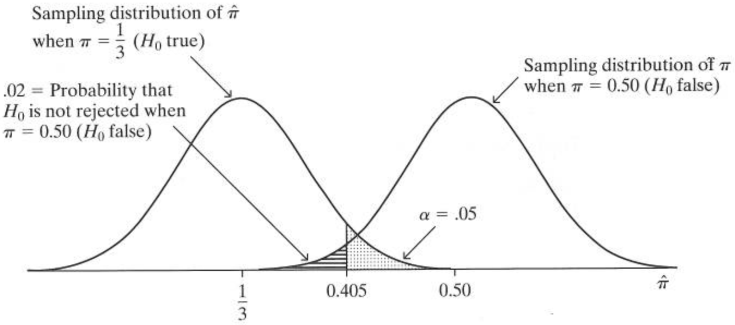
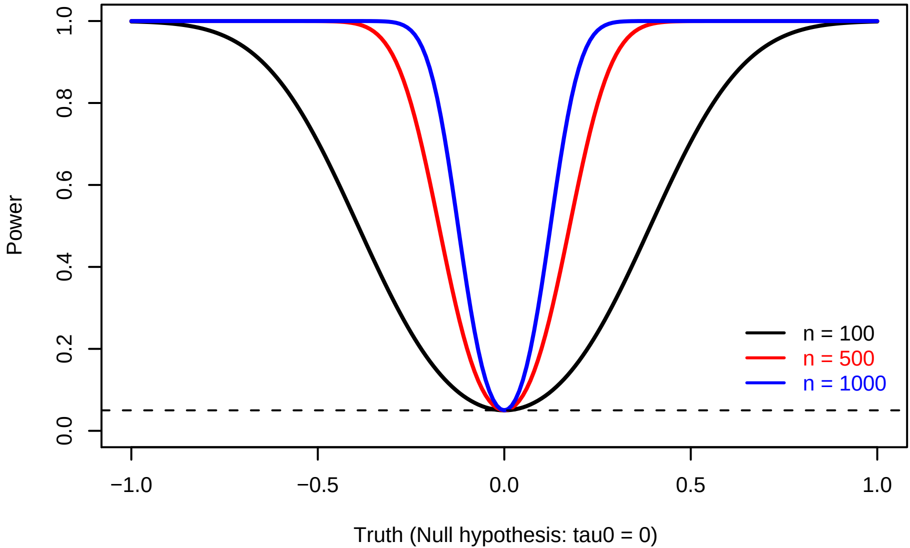



**FIGURE 6.11:** Calculation of $P$(Type II Error) for Testing $H_0$: $\pi = 1/3$ against $H_a$: $\pi > 1/3$ at $\alpha = 0.05$ Level, when True Proportion is $\pi = 0.50$. A Type II error occurs if $\hat{\pi} < 0.405$, since then $P$-value $>0.05$ even though $H_0$ is false.

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



# Paul's Rival, Mani the Parakeet

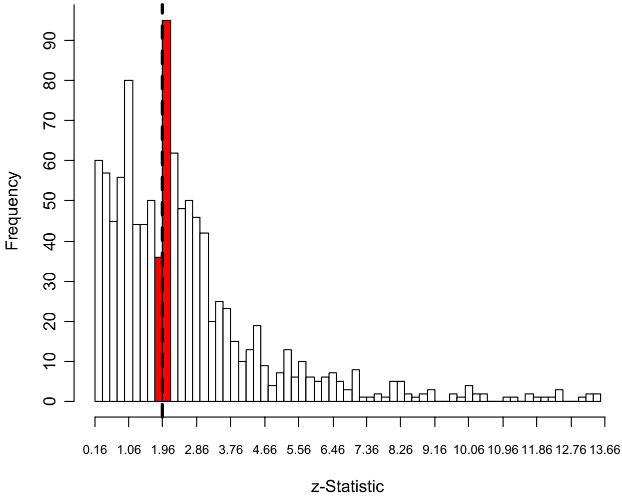
- 2010 World Cup
  - Quarter-final: Netherlands vs Brazil
  - Quarter-final: Uruguay vs Ghana
  - Quarter-final: Argentina vs Germany
  - Quarter-final: Paraguay vs Spain
  - ~~Semi-final~~: Uruguay vs Netherlands
  - Semi-final: Germany vs Spain
  - ~~Final~~: Netherlands vs Spain

- Mani did pretty good too: *p*-value is 0.0625
- Danger of multiple testing $\Longrightarrow$ 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




Gerber and Malhotra, *QJPS* 2008

# Statistical Control of False Discovery

- Pre-registration system: reduces dishonesty but cannot eliminate multiple testing problem
- Family-wise error rate (FWER): $\Pr(\text{making at least one Type I error})$
- Bonferroni procedure: reject the $j$th null hypothesis $H_j$ if $p_j < \frac{\alpha}{m}$ where $m$ is the total number of tests
- Very conservative: some improvements by Holm and Hochberg

- False discovery rate (FDR):

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

- Adaptive: # of false positives relative to the total # of rejections
- Benjamini-Hochberg procedure:
  1. Order $p$-values $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$
  2. Find the largest $i$ such that $p_{(i)} \leq \alpha i/m$ and call it $k$
  3. Reject all $H_{(i)}$ for $i = 1, 2, \ldots, k$

# Key Points

- Stochastic proof by contradiction
  1. Assume what you want to disprove (null hypothesis)
  2. Derive the reference distribution of test statistic
  3. Compare the observed value with the reference distribution
- Interpretation of hypothesis test
  1. Statistical significance $\neq$ scientific significance
  2. Pay attention to effect size
- Power analysis
  1. Failure to reject null $\neq$ null is true
  2. Power analysis essential at a planning stage
- Danger of multiple testing
  1. Family-wise error rate, false discovery rate
  2. Statistical control of false discovery