Analyzing Small Sample Experimental Data

Session 1 - Part 1: Basics

Dominik Duell (University of Essex)

July 15, 2017

Part I: Basics

Introduction

Source: https://mathwithbaddrawings.com/

What's the small sample problem?

- Not enough information!
  - Characteristics of the sample distribution uncertain
  - Inferences about population very problematic
- Harder to establish robust finding
  - Asymptotic properties of tests/estimators not applicable in small samples
  - usually less statistical power

Why do we still analyze small samples?

- Sometimes we cannot or should not collect more data
- Virtually all samples are too small – do not only think number of observations but number of subjects in the study, subject pool, number of stimuli, number of outcome variables, etc.
- Smaller samples allow us to investigate many responses on one subject, one respondent in more detail
- We are often left with detecting only large differences – but is this not what matters? – death to p-hacking ...
- Every additional information will reduce uncertainty about the true effect of interest

What's a small sample?

William Sealy Gosset aka **Student**

- **Gosset** was a brewer with Guinness
- Guinness' problem: experimentation with hops and barley turned out wide variation in quality of brew
- Given the small number of data points, how to understand differences based on variation in hops and barley?
- With help of **Karl Pearson**, Gosset tabled the errors for his question about variation in observed means – known as the **t-distribution**

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

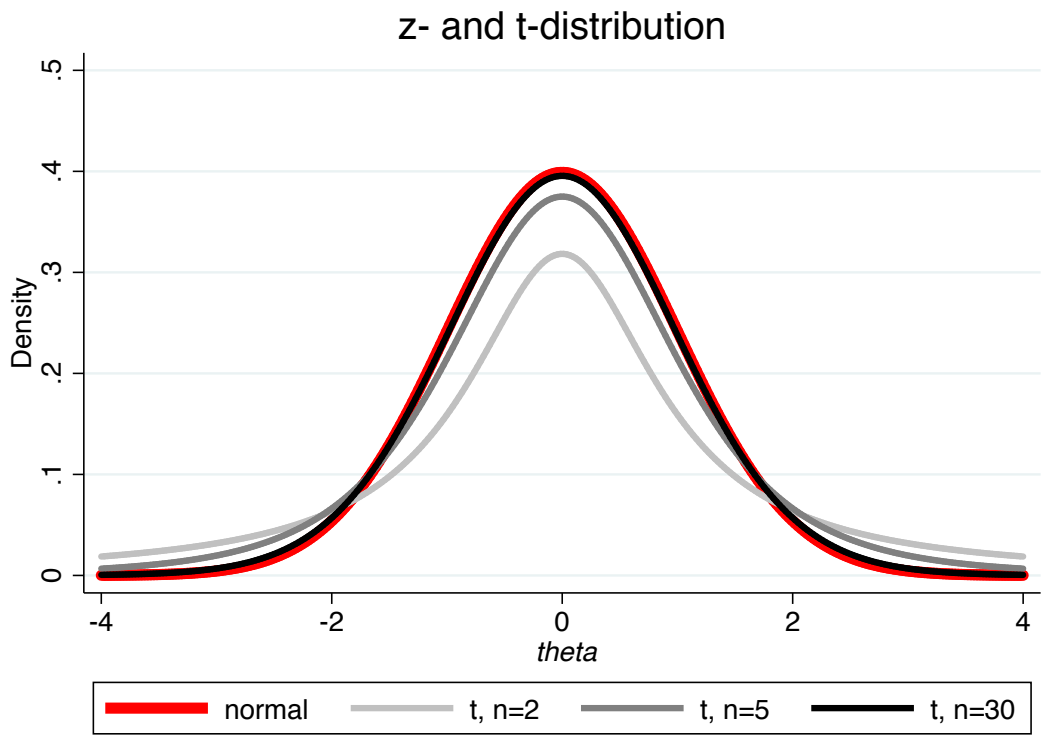
If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in

Biometrika 6 (1), 1908, pp. 1-25

- The central limit theorem says, sampling distribution of a statistic follows normal distribution
- When we know the standard deviation of the underlying population we are able to compute z
- What if sample small and/or we do not know st dev?

$$t = \frac{\overline{\hat{\theta}} - \theta}{s/\sqrt{n}} \tag{1}$$

where $\overline{\hat{\theta}}$ is the sample average of the statistic, s it's standard deviation, n the sample size, and θ is the true parameter of the population



- Small sample: $n \approx 30$?
- Applies when using t-statistic but
  - depends on test/statistic/estimator, population characteristics, sample characteristics
  - better to evaluate small sample properties of test/statistic/estimator, learn all there is about your sample
  - Tests of a particular distribution of a small sample unreliable

Whether you run into small sample issues depends on your sample and the statistical method you are applying

When to think about sample size?

- Research design stage:
  - Budget
  - Randomization
  - Type of measurement
  - Target statistic, target statistical test/estimator
    $\rightarrow$ Power analysis
- Data analysis stage
  - Choosing proper statistical tool
- Research report stage
  - Assumptions
  - Robustness

What do we need?

1. a sense of characteristics of your data
2. given those characteristics, understanding when a small sample runs into problems and which estimator/statistic/test is appropriate ...
   - and relaxes assumptions
     in particular:
     - with respect to outliers
     - or the underlying distribution?
   - and would work with a transformation of the data (quantile, rank)?
   - and tests for what we are actually interested in (i.e., location, distribution, direction)
3. thinking about criteria and ways to evaluate performance of tests/estimators: robustness, size of test, statistical power

We will not talk about how to collect more/better data and research design in detail

Schedule

Saturday morning session 1

We start with a pretty broad introduction to many issues relevant to working with small samples!

- Part I: Basics
  - Exploratory data analysis
  - Fundamentals of analyzing (experimental) data
  - Rubin causal model, treatment effects
  - Analytical frame, exact statistics, randomization inference
  - Standard estimators and sample size
  - Statistical power calculation

Saturday morning session 1

That needs some hands-on practice:

- Part II: Tools and applications
  - Simulations in Stata and R
  - Randomization inference in Stata and R
  - Assessing finite sample properties of standard estimators
  - Computing statistical power

Saturday afternoon session 2

Session 2 and 3 cover a lot of details of various helpful tests and estimators
Non-parametric tests and estimators

- Basics
- Tests for differences between groups
- Alternatives to correlation coefficients
- Confidence intervals

Sunday morning session 3

Non-parametric tests and estimators continued

- Alternatives to parametric regression analysis
- Survival analysis
- More topics in regression analysis:
  - Parametric, semi-parametric, and non-parametric methods
  - Local averaging, local smoothing techniques

Sunday afternoon session 4

Putting it all back together:
Simulations, resampling, and more small sample applications

- Recap simulations
- Bootstrapping
- Small sample applications
  - Small sample properties of parametric and non-parametric estimators
  - Bootstrapping parametric and non-parametric statistics

What could you get out of this course?

- At least a footnote claiming that your results are robust to relaxing assumption X – start a dictionary
- Be more confident whether you can trust your results and the estimator/test you apply
- Get to know your data better – thanks to ideas about simulation, resampling, and visualization but also because statistics and tests you will see make you look at the data from different angles
- Meet the fundamental objective of scientific research: use data to make broader conclusions about the phenomena of interest – uncertain ...

Explanatory data analysis

Why exploratory data analysis?

- Find data entry mistakes
- Check assumptions – but, often cannot tell whether met in small sample
- Center in on appropriate tools for data analysis

How to do exploratory data analysis?

- Summarize:
  - statistics: – mean, median, mode, variance, standard deviation, interquartile range, range, skewness, kurtosis
  - box plots, bar plots, histograms, stem plots, density plots for 1 dimension
  - overlaid 1-D plots, scatterplots for k dimensions
- Guiding questions:
  - Nature of the data? nominal, ordinal, interval
  - Characteristics of distribution? central tendency, dispersion
  - Influential observations? Cooks'D, observation's effect on average, scatter/histogram
  - Representativeness of display

How to do exploratory data analysis – small sample issues

- Print data set!
- median and interquartile range most helpful – remember they are most robust to outliers
- watch out for outliers even more
- Is display misleading? – think about bin size, smoothing parameters

Exploratory data analysis – Examples

- 15 Observations in a **fake dataset** – variables with specific properties – small and larger sample
- Small sample of **General Social Survey Study** 2008-2009

- Fake data

	var	cat	varCorr	varWeakCorr	varInd	varNonMon	varBiRaw	varOutlier	varOutlier~e	varBi	varExp
1	3	0	6.3203351	17.718635	4.2996476	8.75	.03166165	-1	11.278352	0	1.8221188
2	5	0	6.8023788	8.6798996	5.4453523	12.75	.69220938	-1.6666667	15.591281	1	2.7182818
3	3	1	.08865106	.76770747	1.0299267	8.75	.1589295	-1	16.707492	0	1.8221188
4	15	0	12.806618	18.433083	9.9873344	2.75	.12574719	50	68.350896	0	20.085537
5	12	1	11.591313	8.5000767	7.0717553	11	.64991078	-4	-.429595	1	11.023176
6	6	1	1.2054414	1.8113442	3.0015701	14	.69487886	-2	13.9224	1	3.3201169
7	14	1	18.475363	17.34319	11.915388	6	.36206672	-4.6666667	-4.4815005	0	16.444647
8	2	0	3.6108403	2.4362802	7.128063	6	.68953727	-.66666667	-.57501789	1	1.4918247
9	9	0	9.9270512	18.557397	5.3721975	14.75	.12834828	-3	3.7635363	0	6.0496475
10	11	0	14.020344	14.084645	8.7140087	12.75	.48422562	-3.6666667	9.5276606	0	9.0250135
11	15	1	15.447195	14.447053	11.565671	2.75	.15236293	50	61.923961	0	20.085537
12	1	0	5.2772018	8.5049272	11.382254	2.75	.40188543	-.33333333	4.3497227	0	1.2214028
13	13	1	14.13883	15.123469	10.162375	8.75	.17067147	-4.3333333	9.8655914	0	13.463738
14	13	0	8.8743127	13.891425	12.218342	8.75	.6864638	-4.3333333	3.9437348	1	13.463738
15	2	0	1.1058027	2.5814635	11.179432	6	.61015886	-.66666667	14.289787	1	1.4918247

- GSS data

	id	income08	income12	trust08	trust12
1	1	$3,000 to $3,999	$50,000 to $59,999	Can trust	Cannot trust
2	2	$35,000 to $39,999	$60,000 to $74,999	Can trust	Can trust
3	3	$15,000 to $17,499	Under $1,000	Cannot trust	Cannot trust
4	4	$25,000 to $29,999	$35,000 to $39,999	Cannot trust	Cannot trust
5	5	$22,500 to $24,999	$30,000 to $34,999	Can trust	Cannot trust
6	6	$22,500 to $24,999	$6,000 to $6,999	Cannot trust	Cannot trust
7	7	$150,000 and over	$50,000 to $59,999	Cannot trust	Can trust
8	8	$75,000 to $89,999	$35,000 to $39,999	Can trust	Cannot trust
9	9	$90,000 to $109,999	$90,000 to $109,999	Cannot trust	Cannot trust
10	10	$60,000 to $74,999	$75,000 to $89,999	Cannot trust	Cannot trust
11	11	$4,000 to $4,999	$10,000 to $12,499	Cannot trust	Cannot trust
12	12	$30,000 to $34,999	$20,000 to $22,499	Can trust	Cannot trust
13	13	$75,000 to $89,999	$17,500 to $19,999	Cannot trust	Cannot trust
14	14	$50,000 to $59,999	$25,000 to $29,999	Cannot trust	Cannot trust
15	15	$60,000 to $74,999	$75,000 to $89,999	Can trust	Can trust

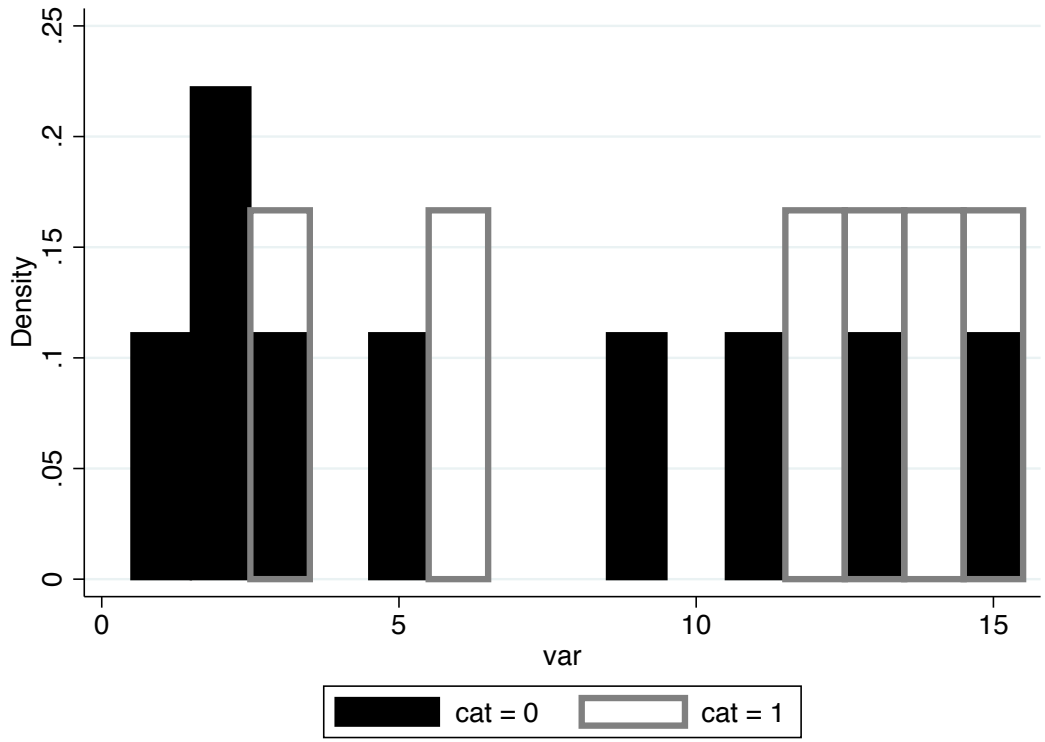
- Sample statistics:
  - `summarize, detail`
  - `histogram var, bin(#)` – for small samples, play with the bin-size and `discrete`-option
  - `stem var` – graphically something from the 90s but also keeps all information
  - `box var` – relies on robust statistics like median and interquartile range

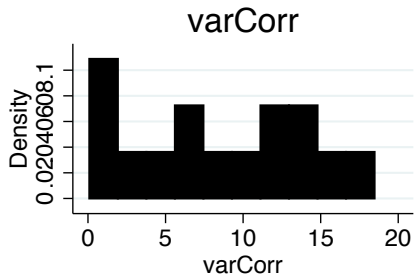
- Fake data

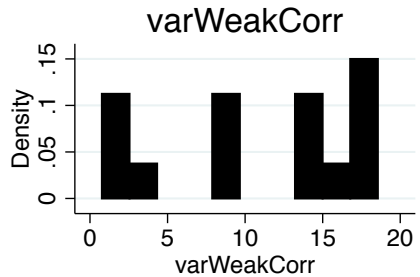
```
                          var
-------------------------------------------------------------
      Percentiles      Smallest
 1%            1             1
 5%            1             2
10%            2             2       Obs                  15
25%            3             3       Sum of Wgt.          15

50%            9                     Mean           8.266667
                      Largest        Std. Dev.      5.297798
75%           13            13
90%           15            14       Variance       28.06667
95%           15            15       Skewness      -.0649552
99%           15            15       Kurtosis       1.359902
```

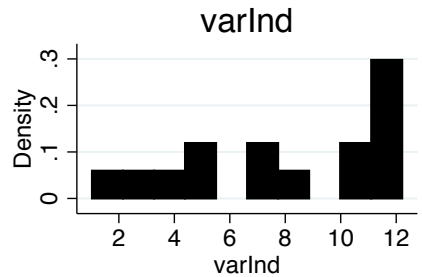
```
        cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          9       60.00       60.00
          1 |          6       40.00      100.00
------------+-----------------------------------
      Total |         15      100.00
```

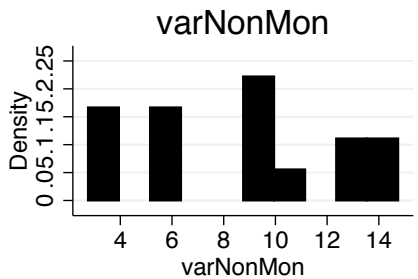
Figure: `hist var, by(cat) bin(10)`

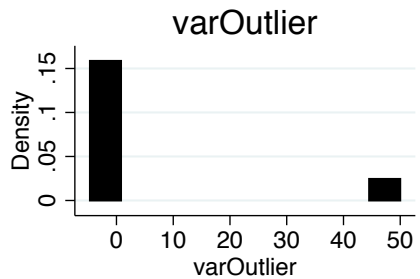


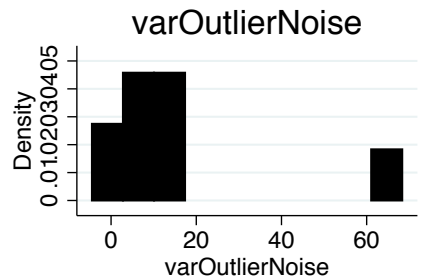

varCorr
Density
0 .02040608.1
0 5 10 15 20
varCorr

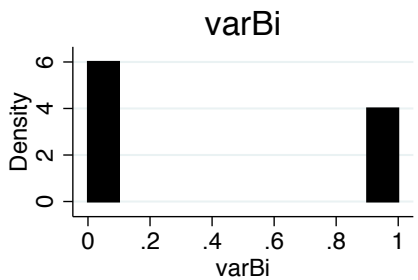

varWeakCorr
Density
0 .05 .1 .15
0 5 10 15 20
varWeakCorr

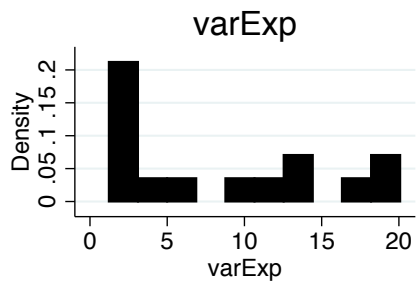

varInd
Density
0 .1 .2 .3
2 4 6 8 10 12
varInd

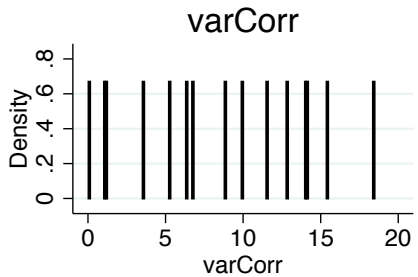

varNonMon
Density
0 .05.1.15.2.25
4 6 8 10 12 14
varNonMon

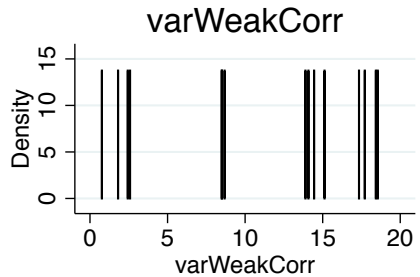

varOutlier
Density
0 .05 .1 .15
0 10 20 30 40 50
varOutlier

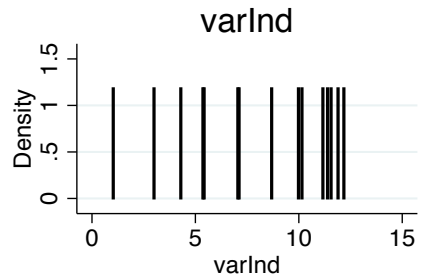

varOutlierNoise
Density
0 .0102030405
0 20 40 60
varOutlierNoise

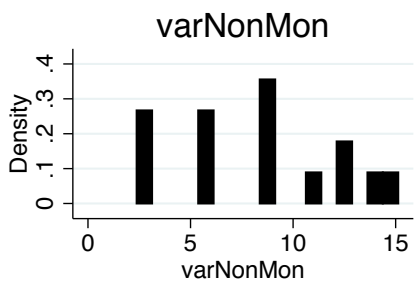

varBi
Density
0 2 4 6
0 .2 .4 .6 .8 1
varBi

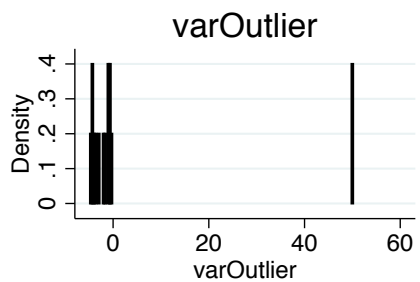

varExp
Density
0 .05.1.15.2
0 5 10 15 20
varExp

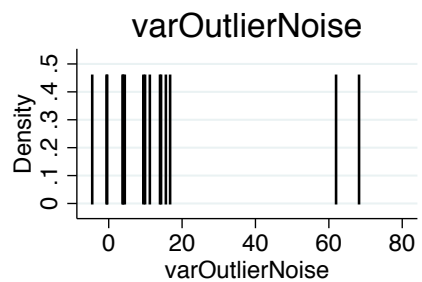

varCorr
Density
0 .2 .4 .6 .8
0 5 10 15 20
varCorr

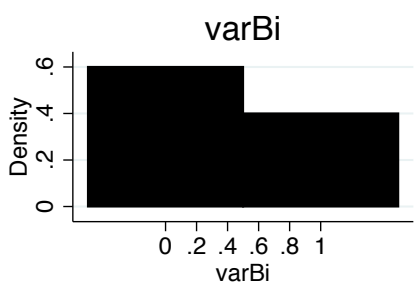

varWeakCorr
Density
0 5 10 15
0 5 10 15 20
varWeakCorr

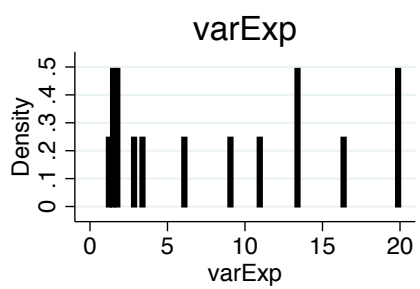

varInd
Density
0 .5 1 1.5
0 5 10 15
varInd


varNonMon
Density
0 .1 .2 .3 .4
0 5 10 15
varNonMon

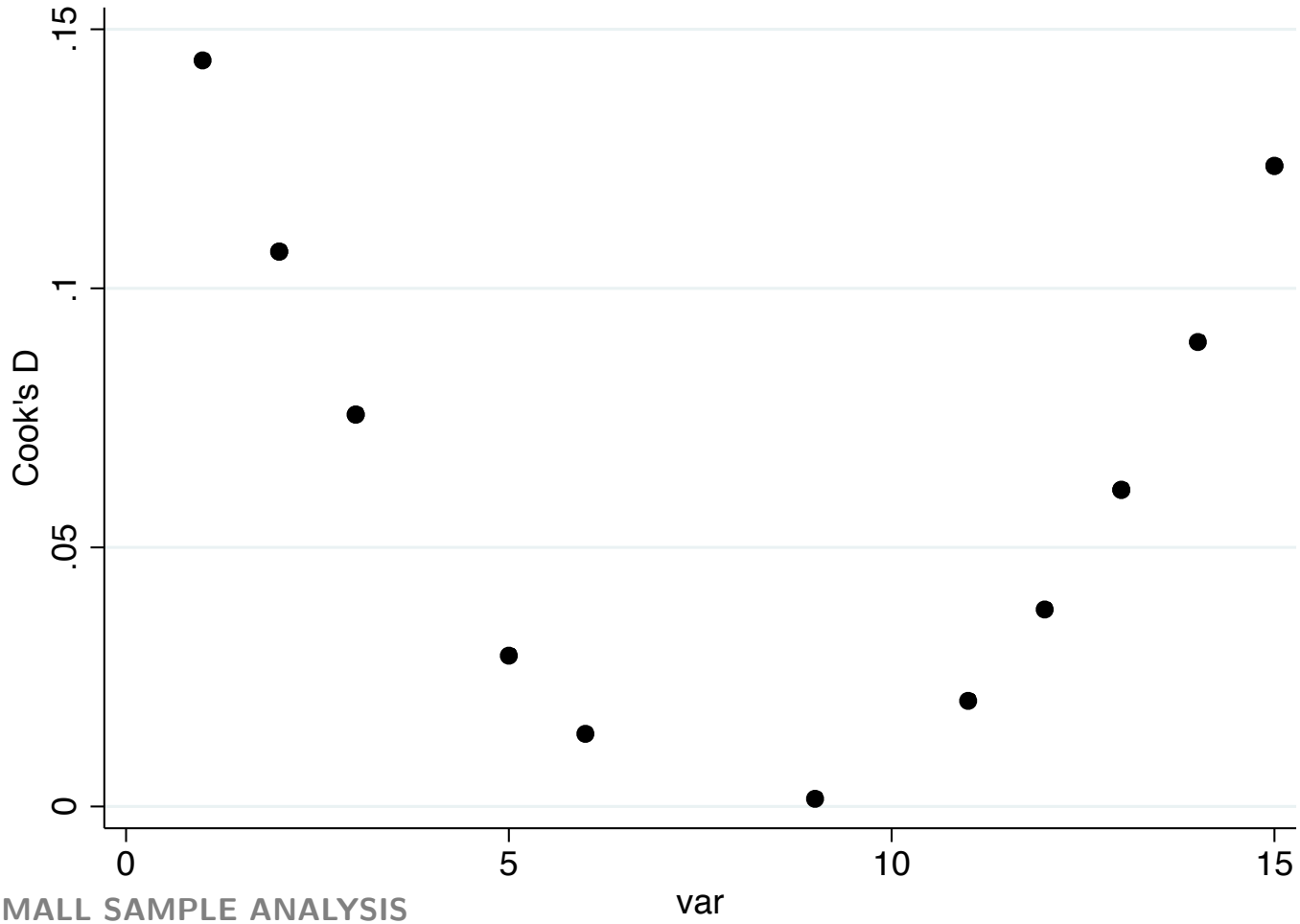

varOutlier
Density
0 .1 .2 .3 .4
0 20 40 60
varOutlier


varOutlierNoise
Density
0 .1 .2 .3 .4 .5
0 20 40 60 80
varOutlierNoise

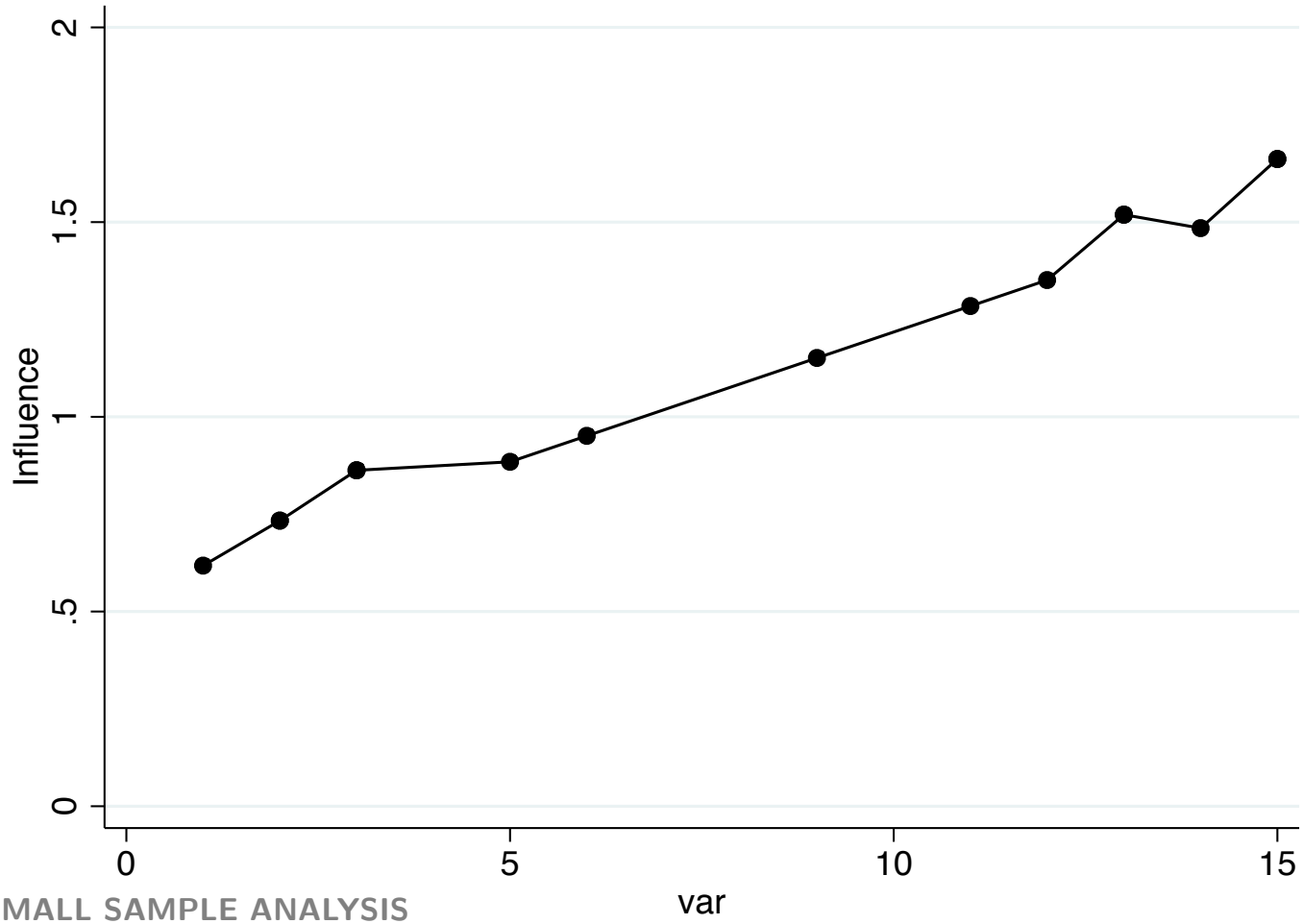

varBi
Density
0 .2 .4 .6
0 .2 .4 .6 .8 1
varBi


varExp
Density
0 .1 .2 .3 .4 .5
0 5 10 15 20
varExp

Influential observations



Influential observations



- GSS data

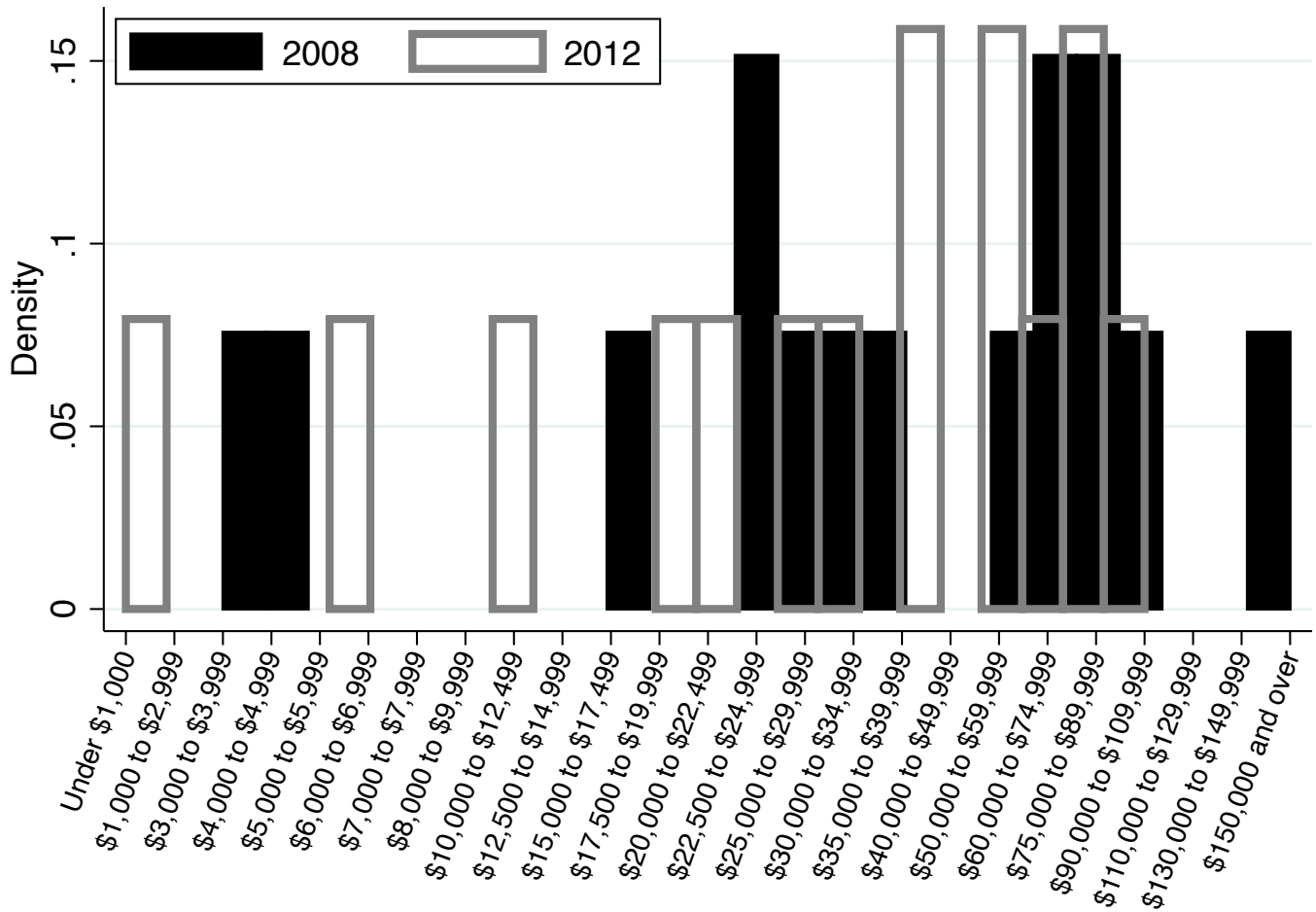
```
. tabulate income08
```

income08	Freq.	Percent	Cum.
$3,000 to $3,999	1	6.67	6.67
$4,000 to $4,999	1	6.67	13.33
$15,000 to $17,499	1	6.67	20.00
$22,500 to $24,999	2	13.33	33.33
$25,000 to $29,999	1	6.67	40.00
$30,000 to $34,999	1	6.67	46.67
$35,000 to $39,999	1	6.67	53.33
$50,000 to $59,999	1	6.67	60.00
$60,000 to $74,999	2	13.33	73.33
$75,000 to $89,999	2	13.33	86.67
$90,000 to $109,999	1	6.67	93.33
$150,000 and over	1	6.67	100.00
Total	15	100.00	

```
. tabulate trust08
```

trust08	Freq.	Percent	Cum.
Can trust	6	40.00	40.00
Cannot trust	8	53.33	93.33
Depends	1	6.67	100.00
Total	15	100.00	

Figure: `gr tw (hist income08, bin(25)) (hist income12, bin(25))`



Fundamentals of analyzing experimental data

You should be somewhat familiar with ...

- probability theory
- random variables and probability distributions
- expected values and parameters of a distribution
- population vs sample, parameter vs estimate
- nominal, ordinal, interval measurement
- discrete, continuous data
- confidence intervals, hypothesis testing

Terminology

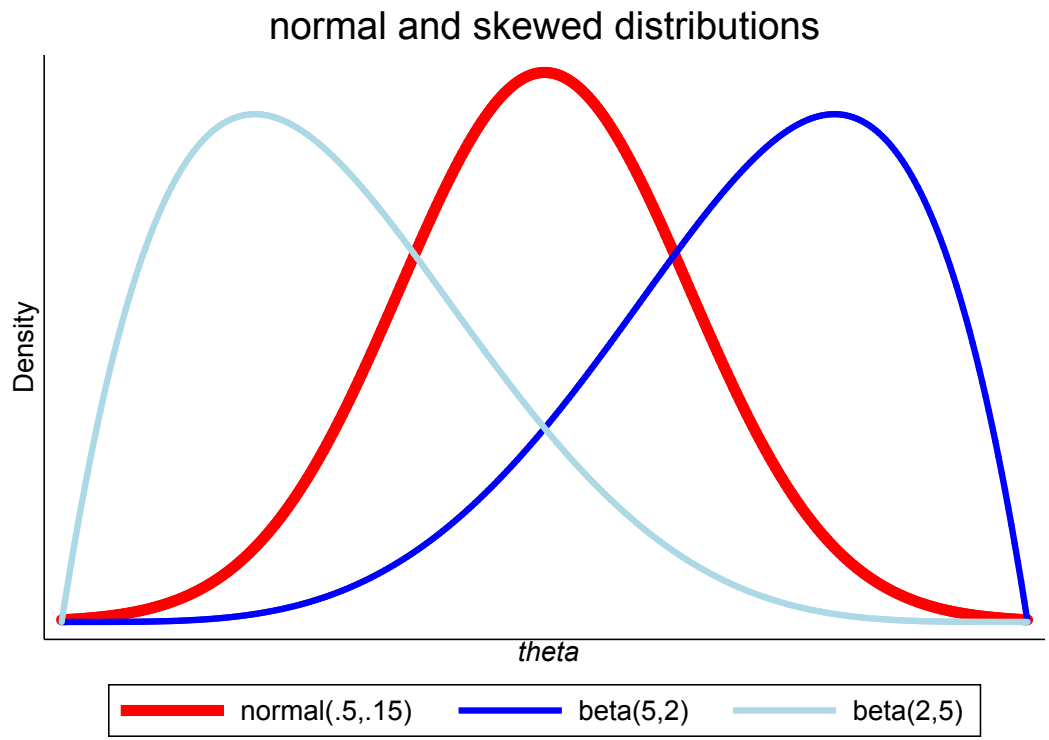
Characteristics of data

- Level of measurement: nominal, ordinal, interval. ratio – rank statistics work with ordinal and interval measurement
- Type of data:
  - categorical: naming/grouping
  - discrete: count
  - continuous: measurement (i.e., interval/ratio)
  - the world is basically discrete, with enough different values we usually assume continuity
  - with continuously measurable data, inference from smaller samples better than with discrete data
- characteristic of function: monotonic

Type of distributions

- normal
- skewed (positive/negative)
- many unnamed

Type of distributions



Normal, t, and F distribution

- t-distribution: family of distributions with degree of freedom (df) as parameter
- F-distribution: family of distributions with numerator df (between-group variability) and denominator df (within-group variability)
- t-test and (Anova) F test specifically designed for small samples – this is somewhat old school, we will look at more recent, more appropriate tests

Potential outcome framework

- Recall: Rubin causal model, you have seen this this week already!
- We obtain our information about the world from data generated by some process – call this the **data generating process (DGP)**
- Sometimes the DGP is given, sometimes researchers manipulate it (i.e, experiment)
- **Experiment:** when a researcher intervenes in the DGP by purposely manipulating elements of the DGP
- **Experimental data:** Data generated by nature and the intervention of an experimentalist
- **Observational data:** Data generated by nature without intervention from an experimentalist

- Assume a population of N individuals, $i = 1, \ldots, N$
- We observe some outcome Y_i for each of them
- There is a treatment variable (independent variable), which either occurs for someone (X=Treatment or $X = T$) or does not (X=Control or $X = C$) – Important: X need not be binary, binary here to introduce the model
- Gives two **potential outcomes** for i:
  Y_i^T and Y_i^C
- Effect of variable X on Y is:
  $Y_i^T - Y_i^C$

That's the **fundamental problem of causality**:

- We do not observe both, Y_i^T and Y_i^C for each individual i
  - We observe Y_i^T if i is in the treatment group (i.e., $X_i = T$)
  - We observe Y_i^T if i is in the control group (i.e., $X_i = C$)
- Any causal claim involves making an assumption about the unobserved **counterfactual**!

How to make proper comparisons between **Treatment** and **Control** in **small samples**?

- **Control** for other confounds either by design or in a regression framework

  Needs proper tools for small samples – potential for omitted variable bias remains

How to make proper comparisons between **Treatment** and **Control** in **small samples**?

- **Random assignment**
  - Estimate **average treatment effect**
    That's what most experimentalists do
    Needs proper tools for small samples – we will look at various parametric and non-parametric tests and estimators
  - **Randomization inference**
    looks at data directly observed and inferred through the null hypothesis – applied to various parametric and non-parametric tests and estimators

Use simulations and bootstrapping to assess the robustness and validity of test/estimators

Analytical frame

Just to be clear ...

- Most of what we look at is in the realm of classical (frequentist) statistics: hypothesis tests
- But, we take the frequentist approach seriously, sample and re-sample: simulations and bootstrap
- I take the p-value associated with a test not to be a decision rule but as additional evidence for a claim also discussed in a broader scientific community (Fisher)
- We also account for small sample issues through more appropriate (non-parametric) test, exact p-values, randomization inference

- Let's fix terms: **statistic** is any function computed from data in the sample – the behavior of a statistic varies with sample size
- Statistics have a **sample distributions** and a certain level of sample variability
- **Point estimate** is a statistic computed from a sample
- Quality of estimates usually assessed by bias (or mean square error or similar) – with small samples those measures become useless
- **Interval estimate** is a range of values containing the true parameter with a certain probability
- **Hypothesis testing**

What's hypothesis testing?

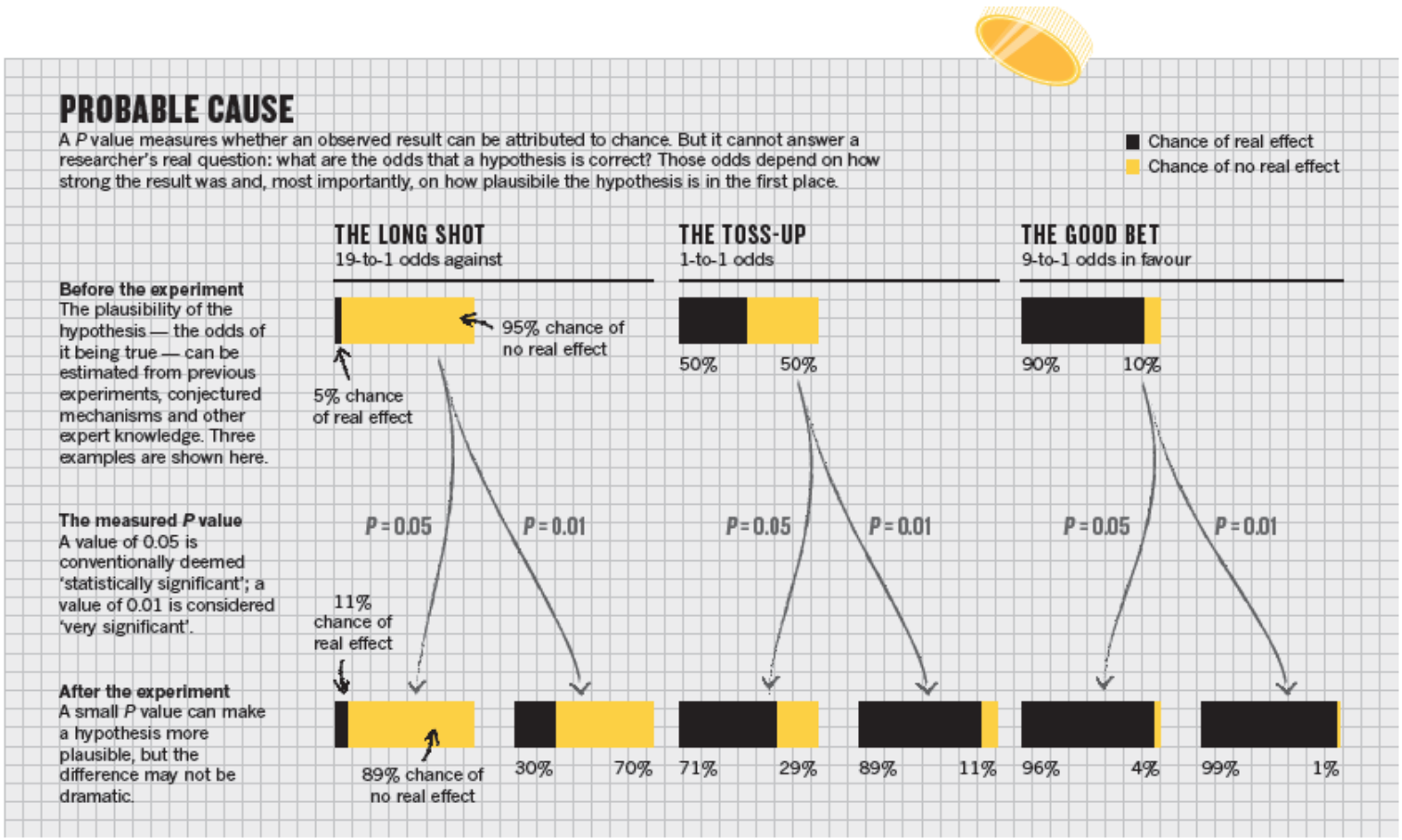
- Rule out chance as explanation for observed effect
- Set up null hypothesis H_0 and reject once enough evidence a deviation from H_0 could not have occurred by chance
- Two kinds of errors:
  - Type I error: Test may lead to rejection of H_0 when it is true
  - Type II error: Test may fail to reject H_0 when it is false
- Size: probability of type I error
- Power: probability that test will correctly lead to rejection of false H_0

Which hypothesis to test?

- Location or dispersion
  - Difference in means]/quantiles or difference in distribution
- Transformations
  - ranks
  - logs
  - ...

  accounting for skewed distributions
- Guided by theoretically reasonable alternative Hypotheses

A cautionary note on p-values



Source: Nuzzo (2015), p.2

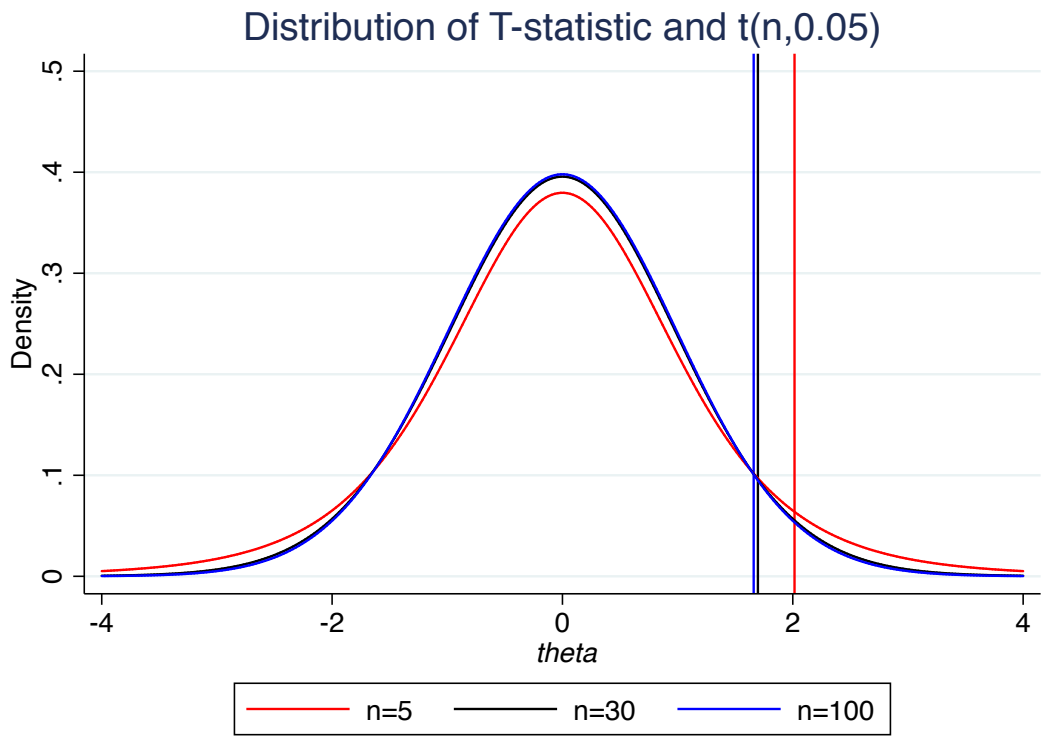
A cautionary note on p-values

- we need to know the prior odds of an effect – independent of your sample size!
- the more implausible the original hypothesis, the higher the probability of a type I error – independent of p-value

Significance and small samples

- Does a high levels of significance imply stronger support for H_a with small sample size than at the same level with larger samples?
  - Yes, because it needs more evidence to reject H_0 in small sample at low p-values (Royall 1986)

Significance and small samples



Significance and small samples

- Does a high levels of significance imply stronger support for H_a with small sample size than at the same level with larger samples?
  - Yes, because it needs more evidence to reject H_0 in small sample at high p-values (Royall 1986)
  - Not immediately: p-value should be understood as the proportion of samples of a given size not providing evidence for H_0 (given that the distribution under H_0 is correct; Knaub 1987)
- With large samples, differences between groups can easily produce small p-values – need effect size to judge whether we care – which is not a problem with small samples

Randomization inference and exact p-values

- Consider the standard hypothesis testing procedure:
  - Derive test-statistics
  - Compare to critical value of a **theoretical** frequency distribution of the test statistic under H_0
  - When observed value is an extreme value in that distribution, H_0 is to be rejected
  - Distribution usually approximates normal (or χ^2-distribution) with increasing sample size
  - Not often appropriate for small samples

Fisher (1935)'s randomization inference:

- Derive test-statistic
- Compare to critical value of distribution of the test statistic H_0 assuming the statistic is a random draw from its randomization distribution
- When observed value is an **extreme** value in that distribution, H_0 is to be rejected
- Randomization distribution is based on all possible combinations of treatment assignment

i	Potential outcomes $Y_i(1)$	$Y_i(0)$	T_i	W^{obs}
1		3	0	
2		5	0	
3	3		1	
4		15	0	
5	12		1	
6	6		1	
$\sum^T / N^T$	7	7.67		-.67

- Observed test statistic here is the average treatment effect: W^{obs}

- What's the randomization distribution?
- Pick H_0
  - Say, the sharp H_0 of no treatment effect for any i: $Y_i^1 = Y_i^0$
- Take observed Y_i^T and fill in the missing Y_i^T-values
- Then compute the test-statistic under all possible combinations of treatment assignment
- Compute **exact** p-value based on the distribution of those test-statistics

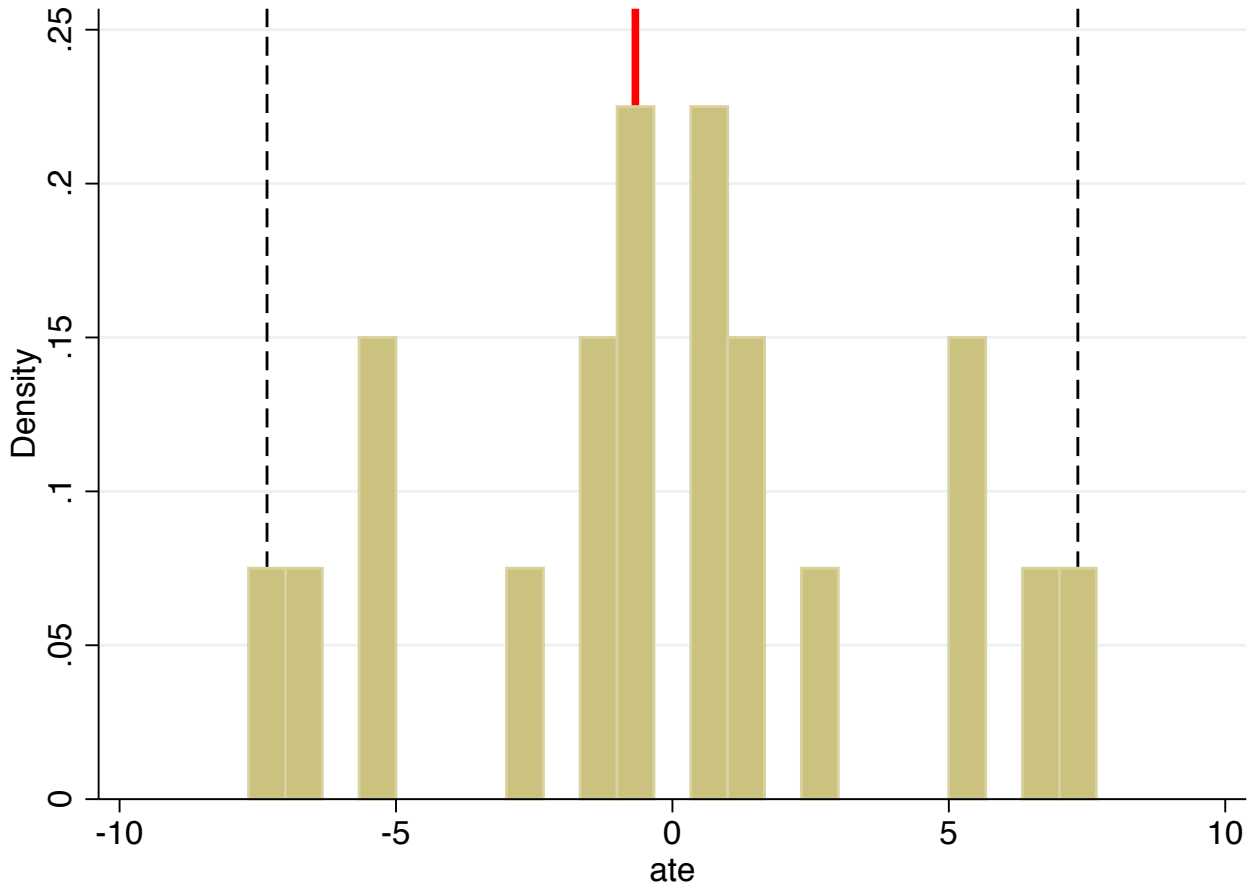
	Potential outcomes			
i	$Y_i(1)$	$Y_i(0)$	T_i	W^{obs}
1	3	3		
2	5	5		
3	3	3		
4	15	15		
5	12	12		
6	6	6		
$\sum^T / N^T$	7	7.67		-.67

- A t-test would give us $p = .55$

	T_1	T_2	T_3	T_4	T_5	T_6	W^{obs}
	0	0	1	0	1	1	-.67

- There are $\binom{6}{3}$ ways to assign 3 of the 6 observations to $T = 1$

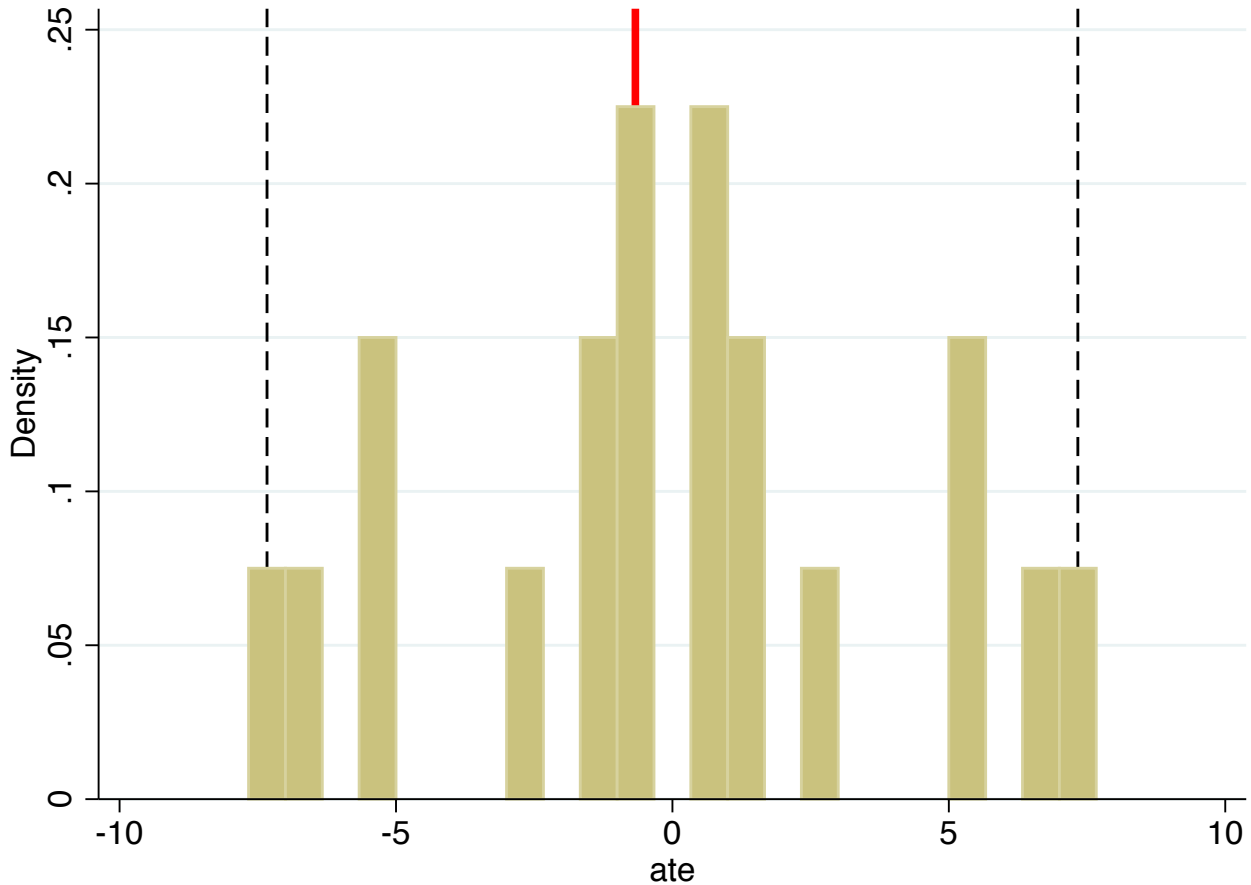
	T_1	T_2	T_3	T_4	T_5	T_6	W
1	0	0	0	1	1	1	7.33
2	0	0	1	0	1	1	-.67
3	0	0	1	1	0	1	1.33
4	0	0	1	1	1	0	5.33
5	0	1	0	0	1	1	-.67
6	0	1	0	1	0	1	1.33
7	0	1	0	1	1	0	5.33
8	0	1	1	0	0	1	-6.67
9	0	1	1	0	1	0	-2.67
10	0	1	1	1	0	0	-.67
11	1	0	0	0	1	1	.67
12	1	0	0	1	0	1	2.67
13	1	0	0	1	1	0	6.67
14	1	0	1	0	0	1	-5.33
15	1	0	1	0	1	0	-1.33
16	1	0	1	1	0	0	.67
17	1	1	0	0	0	1	-5.33
18	1	1	0	0	1	0	-1.33
19	1	1	0	1	0	0	.67
20	1	1	1	0	0	0	-7.33


Density
0
.05
.1
.15
.2
.25
-10
-5
0
5
10
ate

- How to judge whether observed value of the test statistic, W^{obs}, is extreme?
- Under H_0, every possible combination of treatment assignment equally likely to occur
- In our example:
  - $1/20$
  - Measure of extremity: probability of a W at most $-.67$ (lower one-sided test)
  - Let's count

	T_1	T_2	T_3	T_4	T_5	T_6	W
1	0	0	0	1	1	1	7.33
2	0	0	1	0	1	1	-.67
3	0	0	1	1	0	1	1.33
4	0	0	1	1	1	0	5.33
5	0	1	0	0	1	1	-.67
6	0	1	0	1	0	1	1.33
7	0	1	0	1	1	0	5.33
8	0	1	1	0	0	1	-6.67
9	0	1	1	0	1	0	-2.67
10	0	1	1	1	0	0	-.67
11	1	0	0	0	1	1	.67
12	1	0	0	1	0	1	2.67
13	1	0	0	1	1	0	6.67
14	1	0	1	0	0	1	-5.33
15	1	0	1	0	1	0	-1.33
16	1	0	1	1	0	0	.67
17	1	1	0	0	0	1	-5.33
18	1	1	0	0	1	0	-1.33
19	1	1	0	1	0	0	.67
20	1	1	1	0	0	0	-7.33

- How to judge whether observed value of the test statistic, W^{obs}, is extreme?
- Under H_0, every possible combination of treatment assignment equally likely to occur
- In our example:
  - $1/20$
  - Measure of extremity: probability of a W at most $-.67$ (lower one-sided test)
  - Let's count $\rightarrow$ $p = 9/20 = .45$
  - Note that the most extreme p-value possible is $1/20 = .05$ $\rightarrow$ not enough power to find an effect at standard levels of significance


Density
0
.05
.1
.15
.2
.25
-10
-5
0
5
10
ate

- Here is a slightly different way to compute an **exact** p-value
- Now based on all possible combinations of values that yield the observed distribution of outcomes with the given treatment assignment
  - Consider the following contingency table

	Treatment	Control	Total
Outcome 1	2	7	9
Outcome 2	8	2	10
Total	10	9	19

  - Fix marginal frequencies
  - Enumerate all possible contingency tables that produce same marginals
  - Compute probability for each table to judge extremity of observed table

- 10 possible tables with given marginals

	9	0	9
1	1	9	10
	10	9	19

	8	1	9
2	2	8	10
	10	9	19

	7	2	9
3	3	7	10
	10	9	19

	6	3	9
4	4	6	10
	10	9	19

	5	4	9
5	5	5	10
	10	9	19

	4	5	9
6	6	4	10
	10	9	19

	3	6	9
7	7	3	10
	10	9	19

	2	7	9
8	8	2	10
	10	9	19

	1	8	9
9	9	1	10
	10	9	19

	0	9	9
10	10	0	10
	10	9	19

- H_0: no association
- If H_0 is true, how likely would we end up with table 8, 9, or 10
- Compute exact probability that outcome is table 8 or "larger"
- $Prob(outcome) =$

$$\frac{\#\text{ of possibilities favorable to the occurrence of the outcome}}{\text{total }\#\text{ of possibilities}} \tag{2}$$

- 10 possible tables with these marginal totals

	9	0	9
1	1	9	10
	10	9	19

	8	1	9
2	2	8	10
	10	9	19

	7	2	9
3	3	7	10
	10	9	19

	6	3	9
4	4	6	10
	10	9	19

	5	4	9
5	5	5	10
	10	9	19

	4	5	9
6	6	4	10
	10	9	19

	3	6	9
7	7	3	10
	10	9	19

	2	7	9
8	8	2	10
	10	9	19

	1	8	9
9	9	1	10
	10	9	19

	0	9	9
10	10	0	10
	10	9	19

- Total # of possibilities of assignment of values to observations:
  - $\frac{N!}{k!(N-k!)} = \frac{19!}{9!10!} = 92378$ possibilities for each, row and column variable
  - Or $92378 \times 92378 = 8,533,694,884$
- # of possibilities favorable to the occurrence of the outcome:
  - Table 10: exactly one possibility
  - Table 9:
    - $10!/1!9! = 10$ in first column and $9!/8!1! = 9$ in second
    - $10 * 9 * 92378 = 8,314,020$
    - $Prob(\text{Table 9}) = \frac{8,314,020}{8,533,694,884} = \frac{90}{92,378} = .000974258$
  - Table 8: $10!/2!8! = 45$ in first column and $9!/7!2! = 36$ in second
    - $45 * 36 * 92378 = 149,652,360$
    - $Prob(\text{Table 8}) = \frac{149,652,360}{8,533,694,884} = \frac{1620}{92,378}$
  - $\frac{1620+90+1}{92378} = \frac{1711}{92378} = .017536642$

- One-sided test:
- Generally, $Prob(Outcome) = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{N!a!b!c!d!}$
- For

a	b	a+b
c	d	c+d
a+c	b+d	N

- In our example, $Prob(\text{Table 10}) = .000010825$
- $Prob(\text{Table 9}) = .000974258$
- $Prob(\text{Table 8}) = .017536642$
- Sum of those probabilities for a one-sided test is .0185

Two-sided test

- For a two-sided test, compute measure of disproportion $= |\frac{a}{a+b} - \frac{c}{c+d}|$

0.90	**0.69**	0.48	0.27	0.06	0.16	0.37	**0.58**	**0.79**	**1.00**
1	**2**	3	4	5	6	7	**8**	**9**	**10**

- Add outcome probabilities of table 1 and 2
- Two-tailed probability is .023

Where to go from here?

General procedure:

- Define a meaningful hypothesis
- Find a valid and meaningful test-statistic
- Derive the distribution of the test-statistic
  - Theoretical – no your assumptions
  - Exact
  - Simulated
- Judging extremity of observed test-statistic value/likelihood of seeing the observed test-statistic value
- Assess the power of the test-statistic

Statistical Power

Why power calculations?

- To distinguish true effects from noise
- To reduce the probability of a not discovering a true effect – which is in contrast to significance tests that guard against false positives
- To balance finding vs resources
- We assume for now that we are not asking whether we should change the experimental design (i.e., randomization protocol) or outcome measures
- Which sample size is enough to detect a true effect with the statistical method we have in mind

Definitions

Recall,

- Type I error: reject H_0 in favor of H_a when H_0 is true – $\alpha = prob(\text{Type I error})$
- Type II error: fail to reject H_0 when H_a is true – $\beta = prob(\text{Type II error})$
- Then, **Power**: reject H_0 when H_a is true – $1 - \beta$

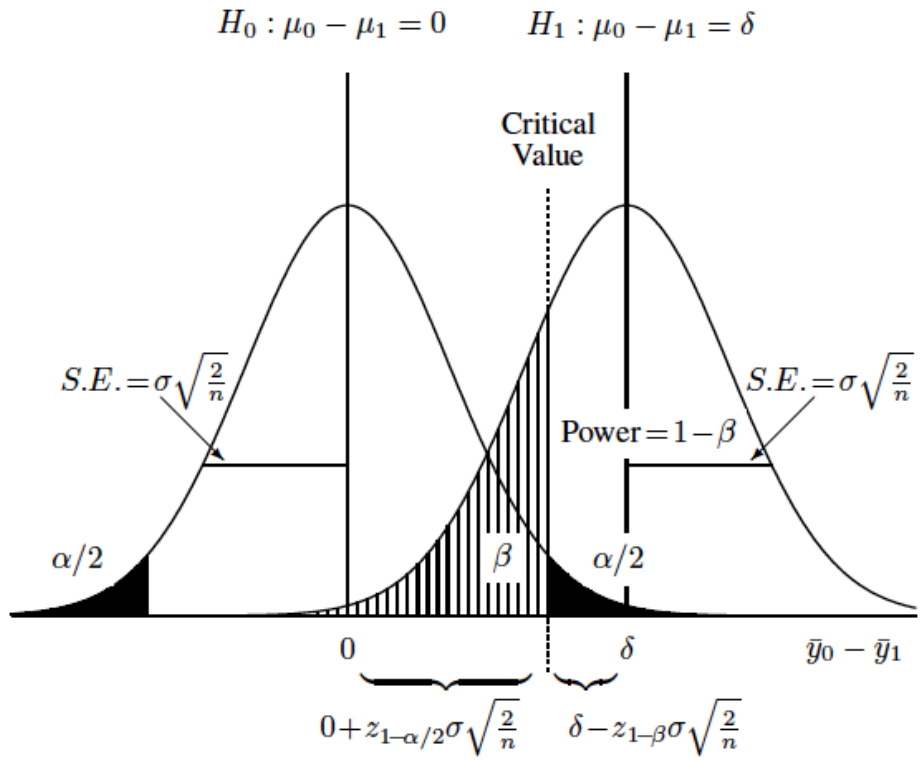


Fig. 2.1 Sampling model for two independent sample case. Two-sided alternative, equal variances under null and alternative hypotheses.

Source: Belle (2008), p.28

References

Basics

- Student (1908): The Probable Error of a Mean
- Fisher (1935): The Logic of Inductive Reasoning
- Royall (1986): The Effect of Sample Size on the Meaning of Significance Tests
- Nuzzo (2014): Statistical Error
- Gelman (2014): Confirmationist and Falsificationist Paradigms of Science
- Morton/Williams (2010): Experimental Political Science and the Study of Causality: From Nature to the Lab

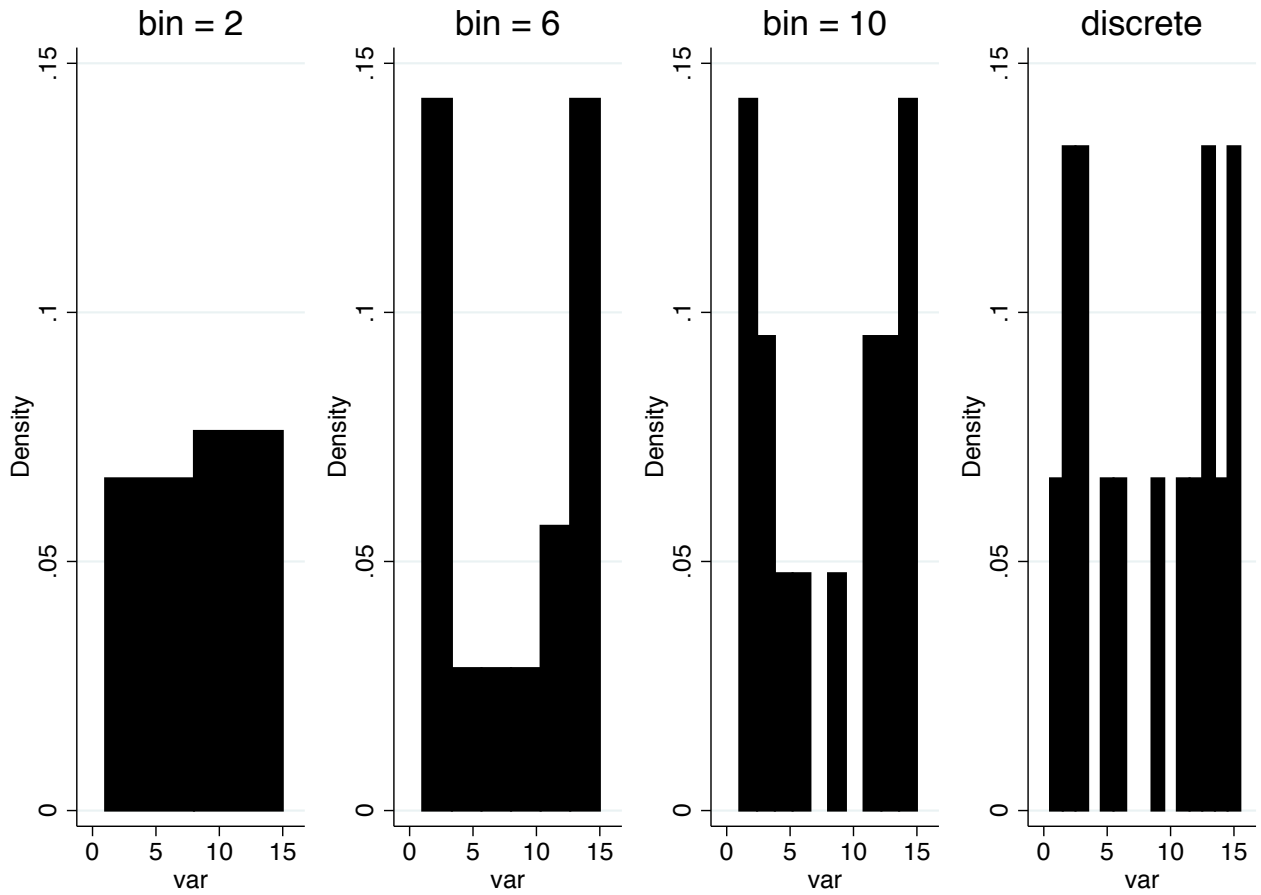
Randomization inference

- Morton/Williams (2010): Experimental Political Science and the Study of Causality: From Nature to the Lab
- Imbens/Rubin (2015): Causal inference in Statistics, Social, and Biomedical Science
- Blackwell (2013, lecture notes): Fishers' Randomization Inference
- Bowers/Panagopoulous (2011): Fisher's randomization mode of statistical inference, then and now.
- Bowers/Panagoulous (2017): Learning from Small Experiments

Power analysis

- EGAP: 10 Things You Need to Know About Statistical Power
- EGAP: Power Analysis Simulations in R

`hist var, bin(#)`



stem var

```
. stem var

Stem-and-leaf plot for var

0* | 1
0t | 2233
0f | 5
0s | 6
0. | 9
1* | 1
1t | 233
1f | 455
```

▸ Go back

```
box var
```