SSSA
Small Sample Statistical Analysis

Day 1
Introduction

Dominik Duell (University of Essex)

September 26, 2016

Source: https://mathwithbaddrawings.com/

What's the problem with small samples?

- Not enough information!
  - Characteristics of the sample distribution uncertain
  - Inferences about population very problematic
- Harder to establish robust finding i.e., 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**

What's a small sample?

- **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**

What's a small sample?

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

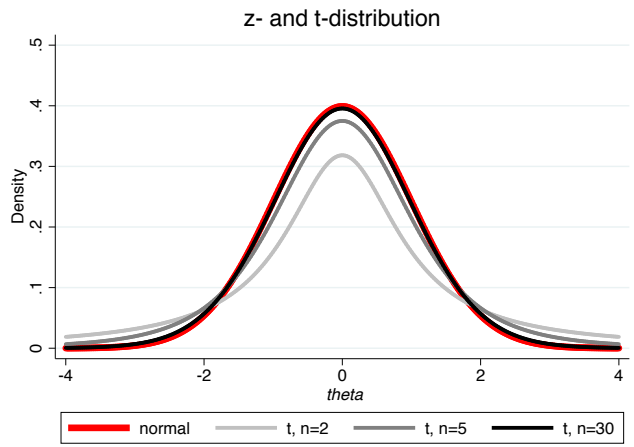
What's a small sample?

- 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

What's a small sample?



What's a small sample?

- Small sample: $n \equiv 30$?
- Applies when using t-statistic but
  - depends on estimator, population characteristics, sample characteristics
  - better to evaluate small sample properties of 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

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 – relax assumptions
3. thinking about statistical power

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

Schedule

Basics - Day 1

- Exploratory data analysis
- Some fundamentals of statistical analysis
- Statistical power calculation

Overview over tools and theory in the morning, basics in running simulations in Stata and R, exercises in simulations and power analysis in the afternoon

Schedule

Non-parametric tests and estimators - Day 2

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

Overview over tools and theory in the morning, exercises in computing and interpreting a selection of non-parametric statistics and confidence intervals, non-parametric statistics in Stata and R in the afternoon

Schedule

Non-parametric tests and estimators continued - Day 3

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

Overview over tools and theory in the morning, exercises in computing and interpreting a selection of non-parametric statistics, non-parametric statistics in Stata and R in the afternoon.

Schedule

Simulations, resampling, Bayesian methods, Multi-level models - Day 4

- Some statistical theory behind Monte Carlo experiments and bootstrapping
- Small sample analysis with simulations and bootstrapping
  - Small sample properties of parametric and non-parametric estimators
  - Bootstrapping parametric and non-parametric statistics
  - Permutation/Randomization test
  - Simulating populations and counterfactuals
- Some Bayesian methods
- Some thoughts on multi-level models
  -

Overview over tools and theory in the morning, exercises in programming simulations and bootstrapping in Stata and R in the afternoon

Schedule

Visualizations - Day 5
In the morning,

- Revisiting exploratory data analysis
- Basic considerations about graphing data
- Thoughts on how many data points are needed to not be mislead by a graphical display
- Simulation and visualization of bias and precision of standard estimators as varying with sample size
- Visually combining simulations and small sample at hand

In the afternoon, get to work - exercise with small sample data set applying (almost) all the tools covered this week

What could you possibly 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 ...

Why?

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

How?

- 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? REMEMBER PROPER MEASURE
  - Representativeness of display

How? – 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

What's in the data

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

Fake small sample

	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

Fake small sample

```
input var cat
        3 0
        5 0
        3 1
        15 0
        12 1
        6 1
        14 1
        2 0
        9 0
        11 0
        15 1
        1 0
        13 1
        13 0
        2 0
```

GSS small sample

	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

Check out your sample

- 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

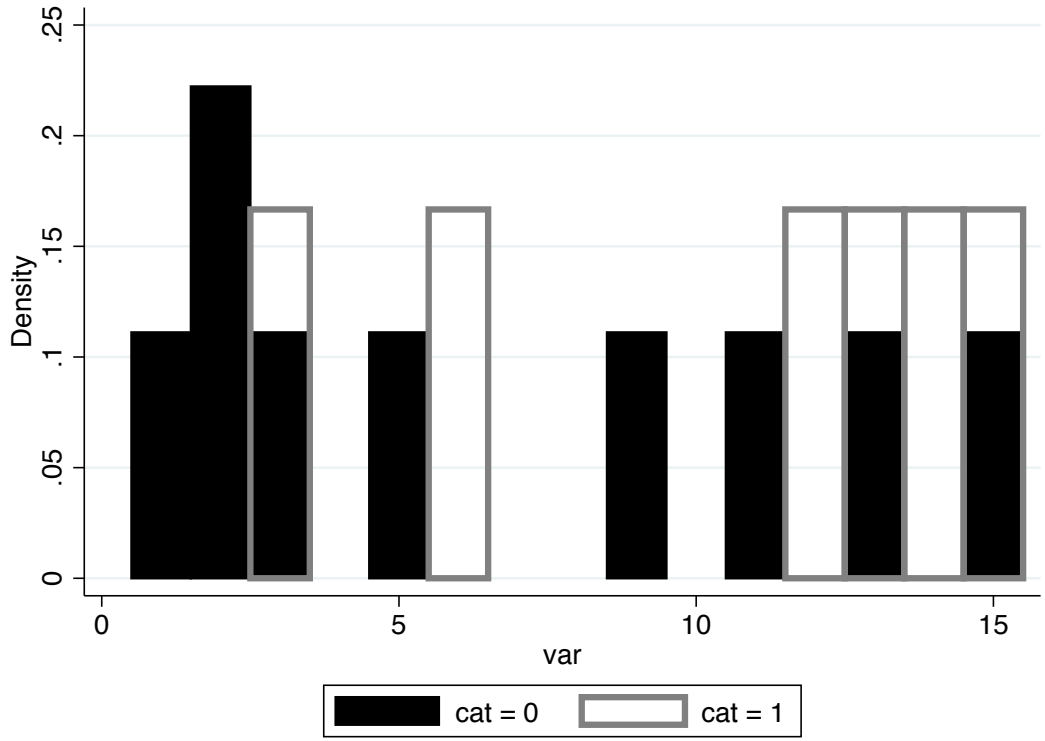
Check out your sample – fake data

```
                          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
```

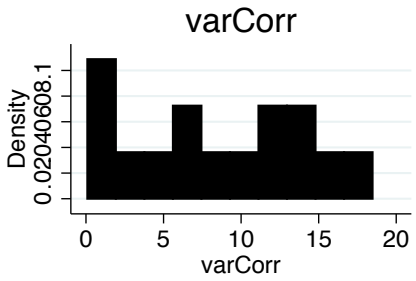
```
        cat |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |          9       60.00       60.00
          1 |          6       40.00      100.00
------------+-----------------------------------
      Total |         15      100.00
```

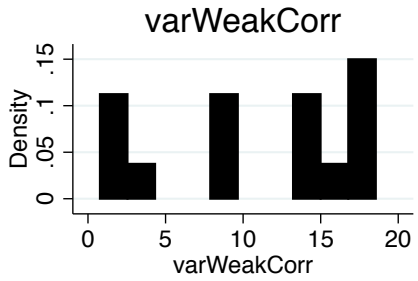
Check out your sample – fake data

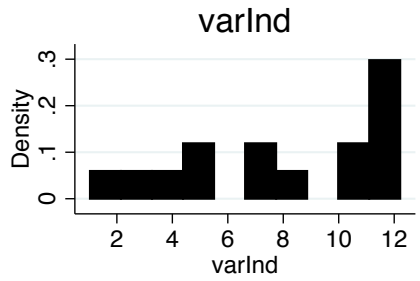
Figure: `hist var, by(cat) bin(10)`

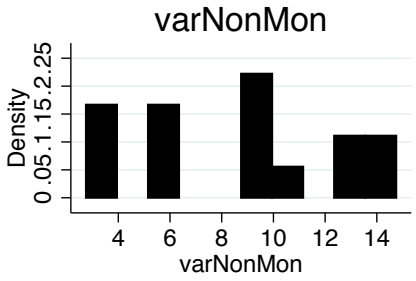


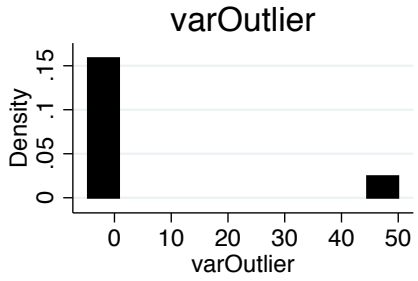
Check out your sample – fake data

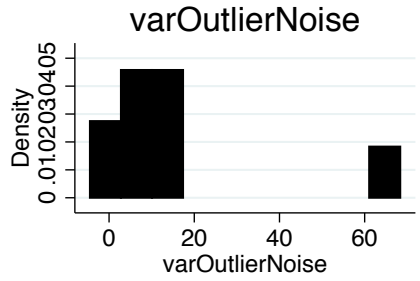


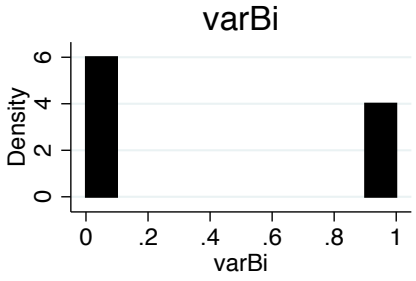


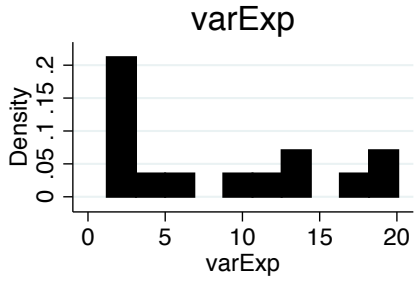




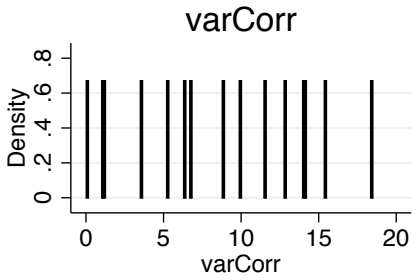


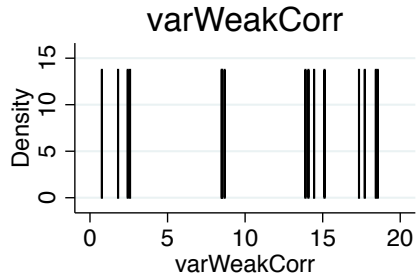


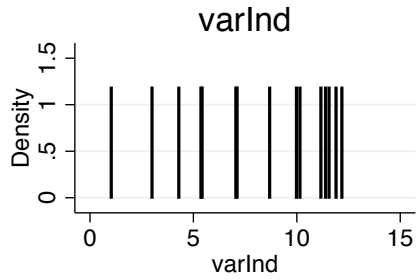


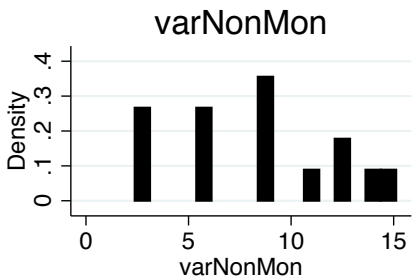


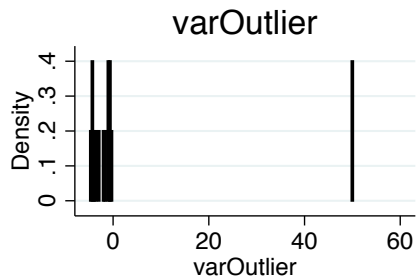
Check out your sample – fake data

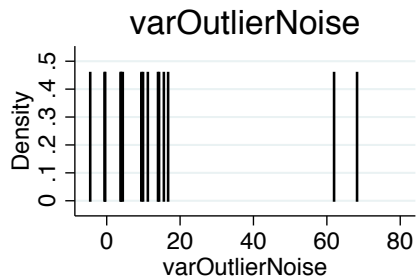


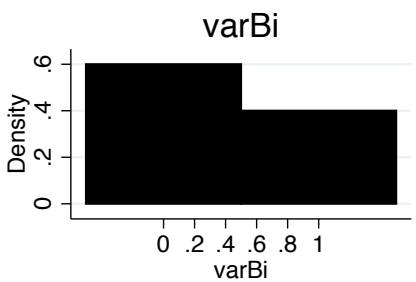


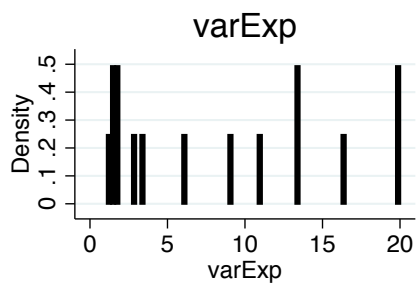










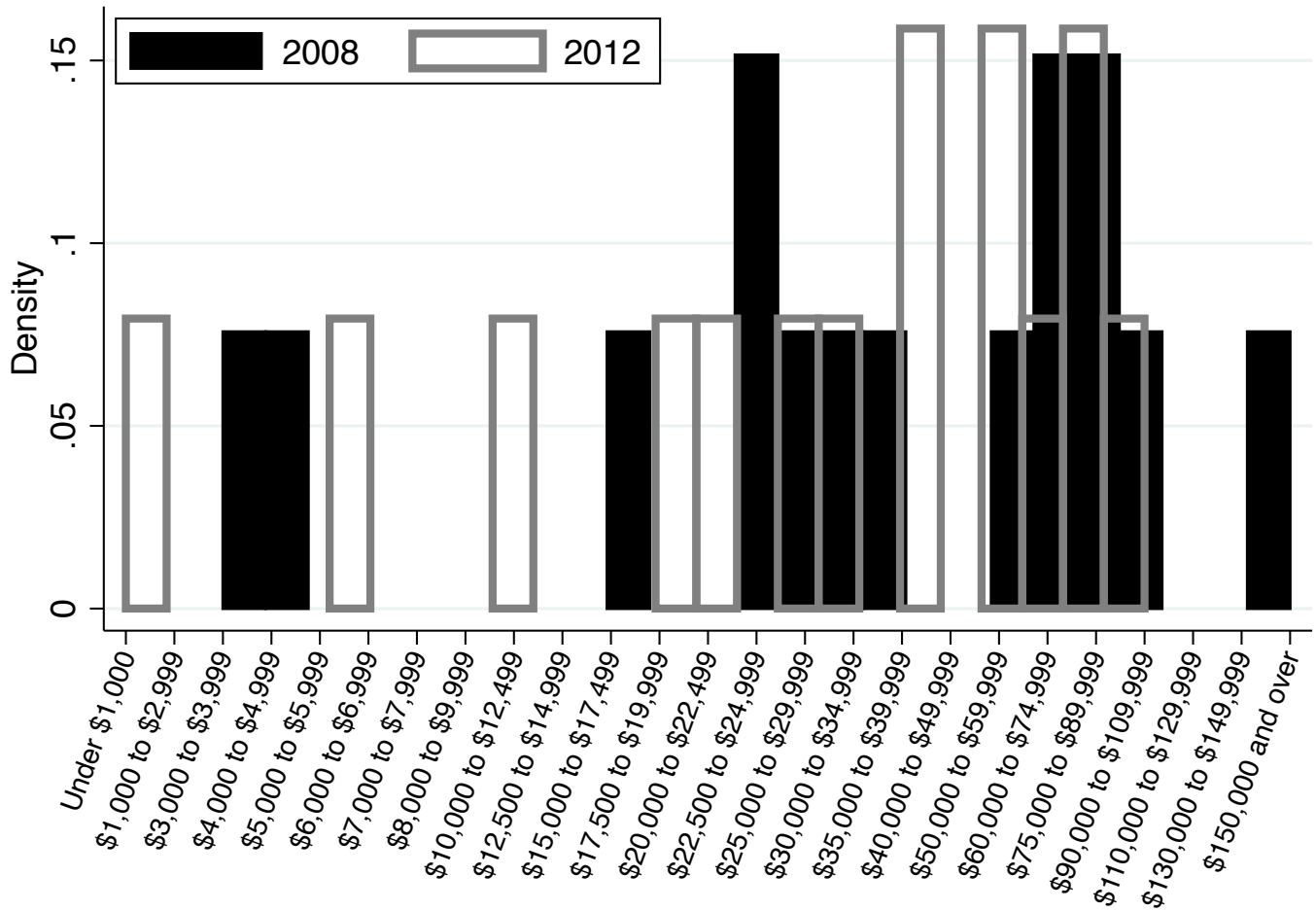


Check out your sample – GSS data

```
. 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
```

Check out your sample – gss data

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



Statistics 101

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

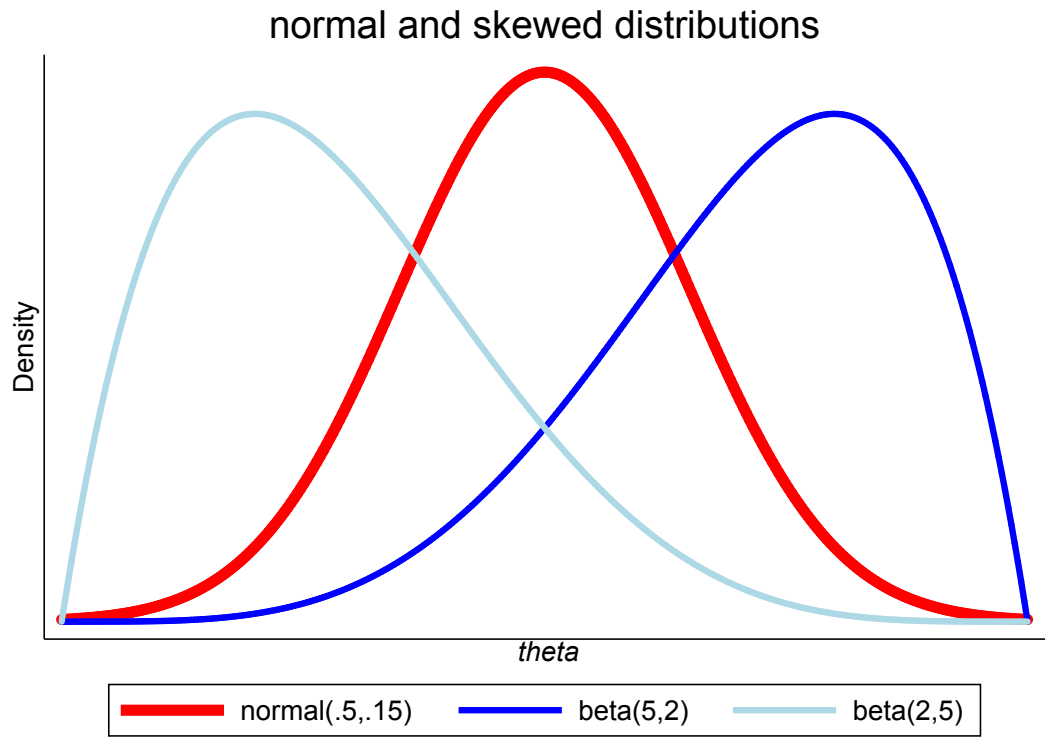
Characteristics of data

- Level of measurement: nominal, ordinal, interval. ratio – rank statistics work with ordinal and interval measurement
- Type of data:
  - categorical: naming/grouping
  - discrete: count
  - continous: 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

Standardized scores

- standard/z-score:
  - to convert normally distributed scores
  - $z = \frac{x-\mu}{\sigma}$
  - where μ is the population mean
  - note, based on assumption about population μ and σ
  - usual rule of thumb, when $n > 30$, sample standard deviation s approximates σ

Standardized scores

- t-score:
  - $t = \frac{\overline{x} - \mu}{s/\sqrt{n}}$
  - where μ is the population mean, $\overline{x}$ the sample mean, s the sample standard deviation, and n the sample size
  - based on assumption about population μ only

Analytical frame

- Most of what we look at is in the realm of classical (frequentist) statistics: non-parametric hypothesis tests
- We take the frequentist approach seriously, sample and re-sample: simulations and bootstrap and then a little excursion into Bayesian statistics – because asserting that any additional information from the smallest sample is valuable implies a Bayesian view

Inferential statistics

- 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**

Hypothesis testing

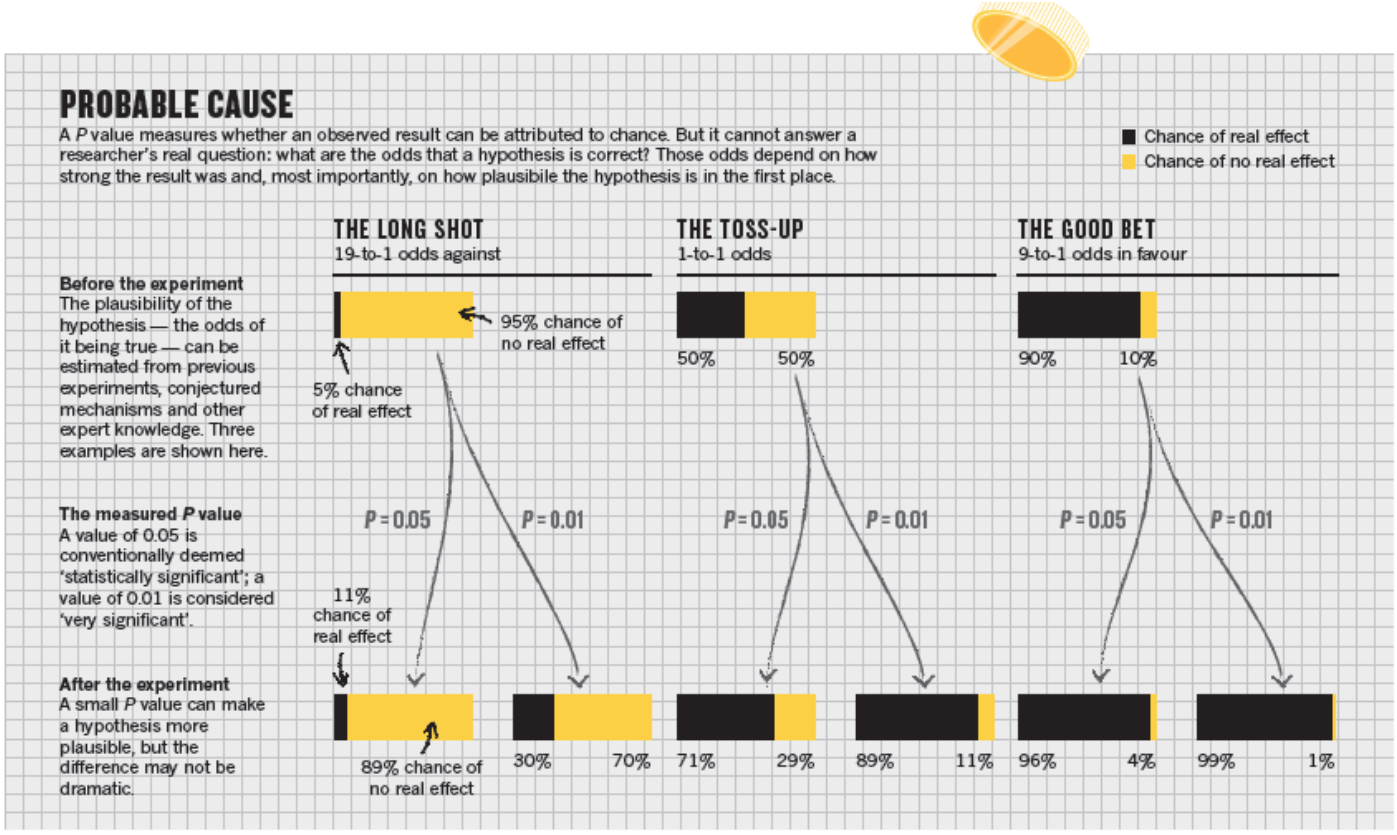
- Rule out chance as explanation for observed effect
- Set up null hypothesis H_0 and reject once enough evidence a deviation from H_0 could not have occurred by chance
- Two kinds of errors:
  - Type I error: Test may lead to rejection of H_0 when it is true
  - Type II error: Test may fail to reject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

Hypothesis testing

- Two ways to think about hypothesis testing:
  - state H_0, when H_0 rejected $\rightarrow$ support for your actual hypothesis of interest H_a ("Confirmationist", see discussion in Gelman 2014) $\rightarrow$ H_a does not need to be stated precisely
  - state H_a, try to find evidence to reject your actual hypothesis of interest ("Falsificationist")

A cautionary note on p-values

- Not meant to be as enshrined as treated today



Source: Nuzzo (2015), p.2

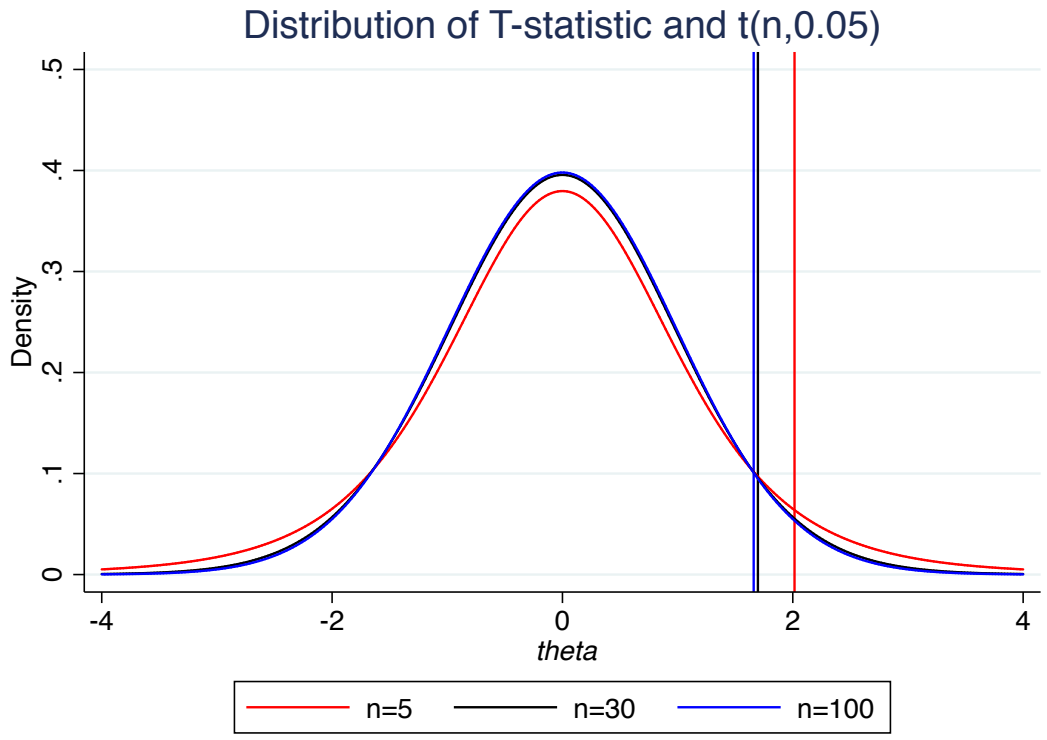
A cautionary note on p-values

- level of significance associated with a p-value of .001 does not say no real effect can only occurs 1% of the time
- 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 high 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

Critical value vs exact probabilities approach

- Critical value vs exact probabilities approach
- Derive test-statistics
- Compare to critical value of a theoretical frequency distribution of the test statistic
  - Distribution usually approximates normal or χ^2-distribution with increasing sample size
  - For small samples, exact probability of obtaining the value of the test statistic has to be derived

Exact statistics

- Based on combinatorial enumeration of all outcomes:
  - Fix marginal frequencies
  - Enumerate all possible contingency tables that produce same marginals
  - Compute probability for each table
- Often permutation tests
- Exact statistics for many estimators/tests available in software packages (e.g. mostly default in R, as option or user-written package in Stata)

Consider the following contingency table

	Treatment	Control	Total
Group 1	2	7	9
Group 2	8	2	15
Total	10	9	19

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

H_0 : *noassociation*

- 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"
- $Prob(outcome) =$

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

Number of possibilities

- Total # of possibilities:
  - $\frac{N!}{k!(N-k!)} = \frac{19!}{9!10!} = 92378$ possibilities for each, row and column variable
  - Or $92378 \times 92378 = 8,533,694,884$
- # of possibilities favorable to the occurence of the outcome:
  - Table 10: exactly one possibility
  - Table 9:
    - $10!/1!9! = 10$ in first column and $9!/8!1! = 9$ in second
    - $10 * 9 * 92378 = 8,314,020$
    - $Prob(\text{Table 9}) = \frac{8,314,020}{8,533,694,884} = \frac{90}{92,378}$
  - Table 8: $10!/2!8! = 45$ in first column and $9!/7!2! = 36$ in second
    - $45 * 36 * 92378 = 149,652,360$
    - $Prob(\text{Table 8}) = \frac{149,652,360}{8,533,694,884} = \frac{1620}{92,378}$
  - $\frac{1620+90+1}{92378} = \frac{1711}{92378} = .0185$

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

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

Statistical power

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 – $1 - \beta$

Statistical power

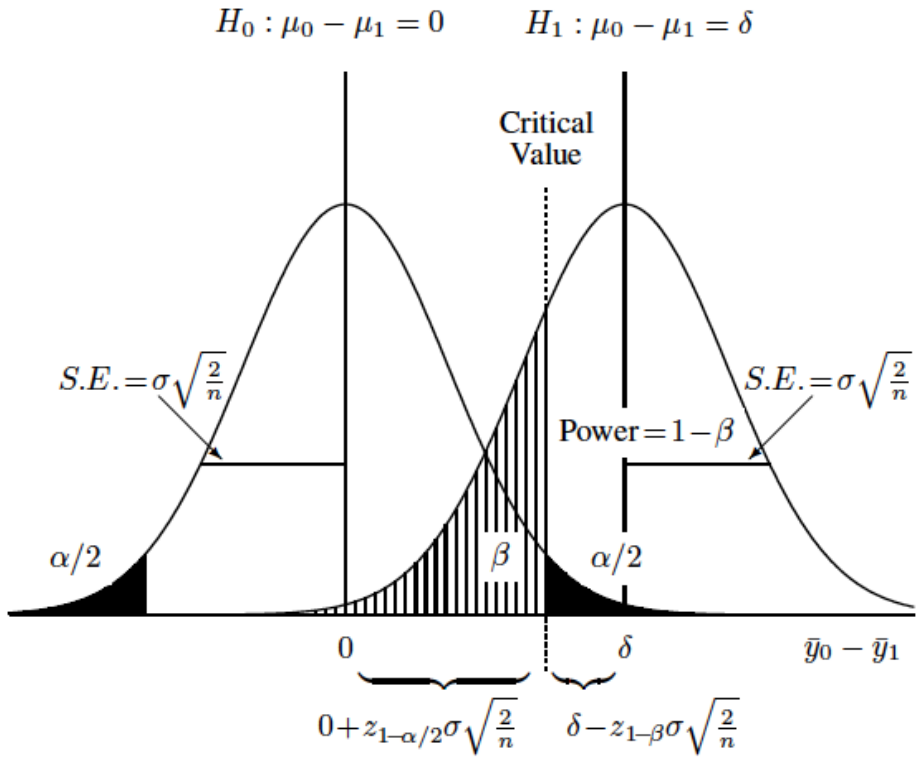


Fig. 2.1 Sampling model for two independent sample case. Two-sided alternative, equal variances under null and alternative hypotheses.

Source: Belle (2008), p.28

Calculating statistical power

Some examples of formulas to compute necessary sample size:

Inference about:

- one sample mean: $N = \frac{z^2_{1-\alpha/2} s^2}{d^2}$ with s anticipated standard deviation and d anticipated bias
- difference in two samples: $N = \frac{(z_{1-\alpha/2}+z_{1-\beta/2})^2(v_1+v_2)^2}{d-(\mu_1-\mu_2)^2}$ with v_i anticipated variance in sample i and μ_i anticipated mean of sample i

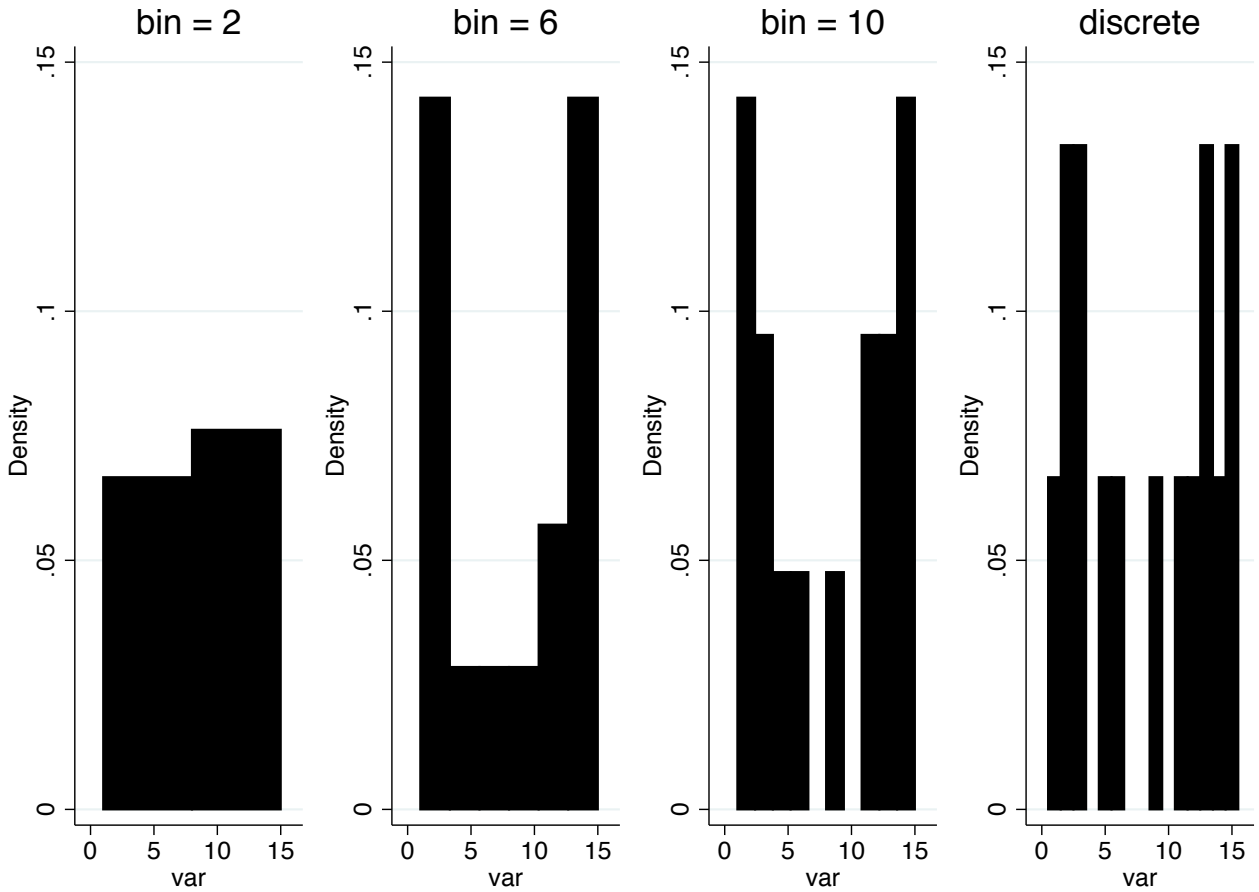
Basics

- Student (1908): The probable error of a mean, Biometrika 6(2), pp.1-25
- Royall (1986): The Effect of Sample Size on the Meaning of Significance Tests, The American Statistician, 40(4), pp. 313-5
- Knaub (1987): Practical Interpretation of Hypothesis Tests, The American Statistician, 41(3), pp. 246-7
- Nuzzo (2014): Statistical error, Nature 506(7487), pp. 150-2
- Gelman (2014): Confirmationist and falsificationist paradigms of science

Power analysis

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

hist var, bin(#)



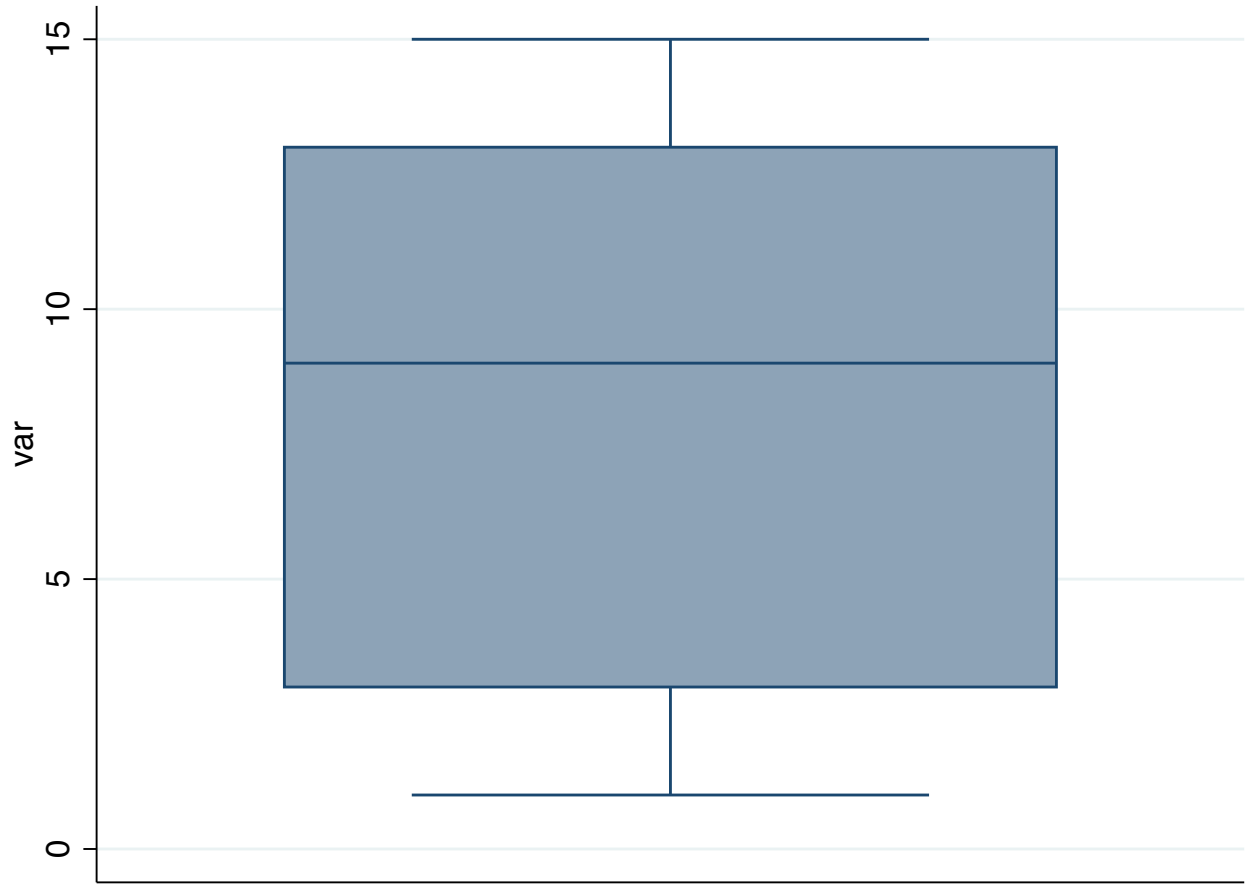
▸ Go back

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



▸ Go back