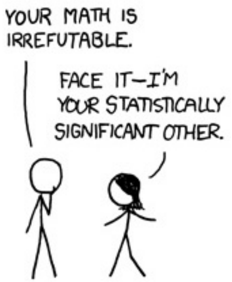
Comparisons



Statistical Reasoning
and Quantitative Methods

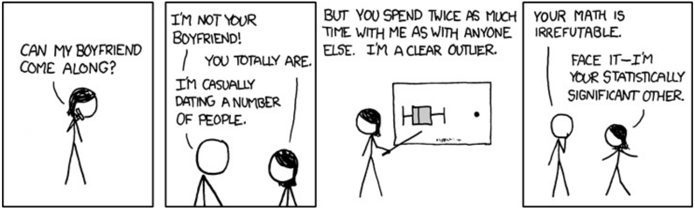
François Briatte & Ivaylo Petev

Session 7

Outline

Significance testing

Comparisons

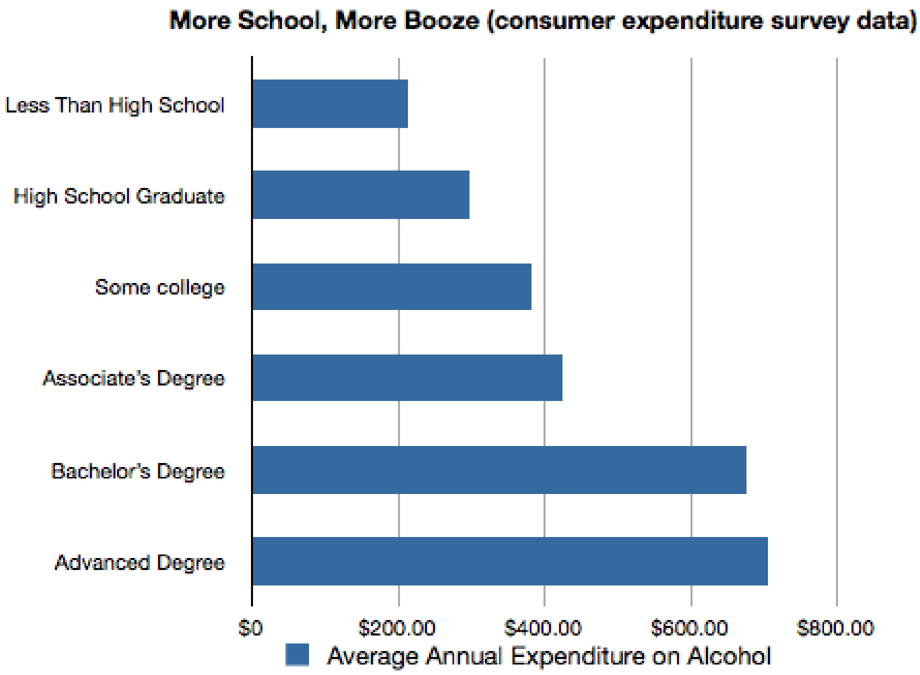


Hypothesis testing

From the Reason Foundation, a "free minds and free markets" U.S. think tank:

> *"A number of theorists assume that drinking has harmful economic effects, but data show that* **drinking and earnings are positively correlated**. **We hypothesize that drinking leads to higher earnings by increasing social capital.** *If drinkers have larger social networks, their earnings should increase. Examining the General Social Survey, we find that self-reported drinkers earn 10-14 percent more than abstainers, which replicates results from other data sets."*
> *(Bethany L. Peters and Edward Stringham, "No Booze? You May Lose", 2006.)*

H_1: "An increase in social drinking causes an increase in earnings."


More School, More Booze (consumer expenditure survey data)
Less Than High School
High School Graduate
Some college
Associate's Degree
Bachelor's Degree
Advanced Degree
$0
$200.00
$400.00
$600.00
$800.00
Average Annual Expenditure on Alcohol

Hypothesis testing

- Devise at least two reasons why the causality might run from high income to frequent social drinking, rather than vice versa. (Question by Cosma Shalizi).
- Imagine a non-linear relationship between alcohol intake, income and employment opportunities in countries where social drinking is legal and generally accepted.

Formally, you first need to **reject the null hypothesis**:
H_0: There is no relationship between social drinking and earnings.
H_a: There is a relationship between social drinking and earnings.

Your alternative hypothesis should be a **directional hypothesis**:
H_{a1} : +social drinking ($\rightarrow$ +social capital) $\rightarrow$ +earnings
H_{a2} : +earnings ($\rightarrow$ +disposable income) $\rightarrow$ +social drinking
...

Significance testing

- **Significance testing** starts with the null hypothesis:
  - H_0 asserts the absence of a relationship.
  - H_a asserts the presence of a relationship.

  Formally, $\Pr(H_0) + \Pr(H_a) = 1$.
- **Proof by contradiction** works by rejecting the null hypothesis:
  - α is a conventional level of statistical significance.
  - H_0 is rejected when its estimated probability is lower than α.

  Conventionally, $\alpha = 0.05$.
- **Serious issues** occur when attaching p-values to hypotheses:
  - Due to formal assumptions, $p = .01 \nRightarrow \Pr(H_a) = .99$.
  - Due to conventions, H_0 is rejected at $p = .051$, retained at $p = .049$.

  H_0 can hence be erroneously rejected or retained.

Significance testing

- **Assumptions** behind significance tests:
  - The population distribution is assumed to be *approximately* normal: $X \sim \mathcal{N}(\mu, \sigma^2)$ by virtue of the Central Limit Theorem.
  - The sample distribution *will* depart from the normal distribution to some extent, by virtue of the Law of Large Numbers.
  - The probability distribution, like Student's t-distribution, reflects only an *estimated* p-value for H_0.
- **Errors** with significance tests:
  - **Type I Error** rejects H_0 when it is actually true.
  - **Type II Error** retains H_0 when it is actually false.

You *cannot* rule out the possibility that your significance test violates its background assumptions, and that you are hence making a Type I or Type II error when interpreting its results, even when $p \ll \alpha$.

Type I and II Errors

- **Type I Error** in judicial trials:

  "Last year executed man proven innocent by DNA evidence."

  - H_0: presumption of innocence
  - H_a: ... until proven guilty (H_0 wrongly rejected)

- **Type II Error** in child protection:

  "Violent father beats children after being released from custody."

  - H_0: parents considered responsible
  - H_a: ... until proven abusive (H_0 wrongly retained)

Proof by contradiction is context-dependent: a Type I Error can carry more serious consequences than a Type II Error, and vice versa.

Medical trials provide some evidence of the risks of both Type I Errors ("tea $\rightarrow$ cancer") and Type II Errors ("therapy $\rightarrow$ death").

Comparison of means in the sample

"Girls suck at math."

Dependent variable: continuous standardised math score (`0-100`)
Independent variable: binary gender groups (`1 "Female" 0 "Male"`)

- **Measurement** in two independent groups:
  - Mean math score for men: $\bar{x}_{men}$ and women: $\bar{x}_{women}$
  - Difference in mean scores: $\Delta_{scores} = \bar{x}_{men} - \bar{x}_{women}$
- **Association** between gender and educational attainment:
  - $H_0 : \Delta_{scores} = 0 \Leftrightarrow \bar{x}_{men} - \bar{x}_{women} = 0$ (no significant difference)
  - $H_a : \Delta_{scores} \neq 0 \Leftrightarrow \bar{x}_{men} - \bar{x}_{women} \neq 0$ (significant difference)
- **Direction** of the alternative hypothesis:
  - $H_{a1} : \Delta_{scores} > 0 \Leftrightarrow \bar{x}_{men} > \bar{x}_{women}$ (men outperform women)
  - $H_{a2} : \Delta_{scores} < 0 \Leftrightarrow \bar{x}_{men} < \bar{x}_{women}$ (women outperform men)

Comparison of means by estimation

"Girls suck at math." $\Leftrightarrow Pr(\Delta_{scores} \neq 0) = 1 - Pr(H_0)$

A **"p-value"** designates the probability level of the null hypothesis.

A significance test rejects H_0 when $Pr(H_0) < \alpha \Leftrightarrow p < \alpha \Leftrightarrow p < .05$ when α, your **level of statistical significance**, is at 95% **confidence** (the conventional standard for our course of study):

- **Reject** H_0 if $Pr(\Delta_{scores} = 0) < 0.05 \Leftrightarrow Pr(H_0) < \alpha$.
  - H_0 is wrongly rejected in *at most* $\alpha = 5\%$ of the cases (Type I Error).
  - H_a is estimated significant *at least* at $1 - \alpha = 95\%$ confidence.
- **Retain** H_0 if $Pr(\Delta_{scores} = 0) > 0.05 \Leftrightarrow Pr(H_0) > \alpha$.
  - H_0 is wrongly retained in $\alpha = 5\%$ of the cases (Type II Error).
  - H_a is estimated insigificant *at least* at $1 - \alpha = 95\%$ confidence.

Comparison of means by statistical significance

"Girls suck at math."
$\Leftrightarrow$ **rejected** *if* $p > .05 \Leftrightarrow Pr(H_0) > \alpha$ *(significance)*
$\Leftrightarrow$ **accepted** *if* $p < .05 \Leftrightarrow Pr(H_0) < 1 - \alpha$ *(confidence)*

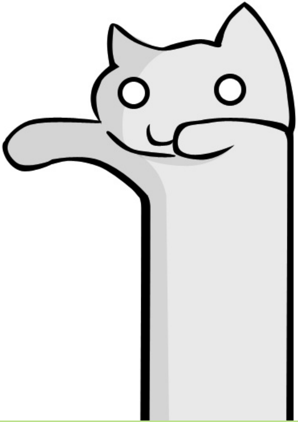
- A significance test **can only reject the null hypothesis** by estimating its *p*-value, which *you* then choose to reject or retain *via* the level of confidence at which you estimate your hypotheses.
- A significance test **cannot prove the alternative hypothesis**, because that interpretation is *your* initiative after reading the estimated probability of the null hypothesis.
- A significance test **cannot attach a p-value to a hypothesis**, since probability levels are estimations based on the null hypothesis. The *"$p < .05$ good, rest bad"* cargo cult is irresponsible.

The Prophecy

The Prophecy requires **raising the power of Estimation Cat** as much as possible:

- **Maximise sample size:** missing observations and/or low *n* reduce the statistical power of your sample.
- **Approach normality:** Estimation Cat is pleased when your variables follow a normal distribution.
- **Use the appropriate test:** significance tests come with restrictive assumptions, e.g. independent groups, equal variance.

This is difficult: Estimation Cat is needy.



The Prophecy

The Prophecy also requires **curbing the threat of Significance Cat** as much as possible:

- **Maximise sample size** (again): the number of observations restricts the degrees of freedom used to calculate p-values from Student's t-distribution.
- **Use the appropriate test** (again): a correct reading of an inappropriate test is a Type III Error (providing the right answer to the wrong question).

This is also difficult, especially given the cunning nature of Significance Cat.

The Prophecy

The Prophecy is chiefly achieved through **statistical reasoning** that carefully balances the awesome powers of estimation and significance:

- **Keep an open mind**: reading p-values require paying attention to Type I and Type II Errors. The key to any test remains interpretation.
- **Avoid cargo cults:** statistical inference is a probabilistic method. Each test result is expressed as a likelihood that carries no absolute certainty.

And yes, this amounts to a lot of computational and intellectual effort for the incomplete and imperfect results of frequentist statistics.

Stata implementation: *t*-test (means)

Compare the average literacy rates of democracies and dictatorships.

```
. bysort democ: su literacy

-------------------------------------------------------------------------------
-> democ = 0. Dictatorship

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    literacy |         74    60.82432    22.97429         18         95

-------------------------------------------------------------------------------
-> democ = 1. Democracy

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    literacy |         96    83.90625    20.19749         21         99

-------------------------------------------------------------------------------
```

Stata implementation: ttest

```
. ttest literacy, by(democ)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
0. Dicta |      74    60.82432    2.670707    22.97429    55.50161    66.14704
1. Democ |      96    83.90625    2.061397    20.19749    79.81386    87.99864
---------+--------------------------------------------------------------------
combined |     170    73.85882    1.861443    24.27025    70.18415     77.5335
---------+--------------------------------------------------------------------
    diff |           -23.08193    3.317918               -29.63211   -16.53174
------------------------------------------------------------------------------
    diff = mean(0. Dicta) - mean(1. Democ)                        t =  -6.9568
Ho: diff = 0                                     degrees of freedom =      168

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

Read the confidence interval of the difference: the interval does not include 0, indicating a statistically significant difference.

Comparison with proportions

"Scotland has the most redheads."

Dependent variable: binary hair colour (`1 "Red" 0 "Other"`)
Independent variable: binary location (`1 "Scotland" 0 "Other"`)

- **The mean of the dependent variable reads as a proportion:** e.g. a mean of $\bar{x} = .13$ indicates a proportion of 13% of redheads.
- **Yet proportions are not statistically assimilable to means:** having red hair is not a continuous but a binary yes/no attribute.
- **A different distribution therefore applies to binary variables:** the binomial distribution is used instead of the normal distribution. The rest of the mechanics are identical:
  - The proportion p of N observations provides a standard error.
  - The difference in proportions Δ_{p-q} provides a confidence interval.
  - The margin of error is half the width of the confidence interval.

Stata implementation: proportions

Do parliamentary regimes select more female leaders?

```
. tab leader parliamentary, col nokey
```

Female leader	Parliamentary Republic 0	1	Total
0. Male leader	124 96.12	37 92.50	161 95.27
1. Female leader	5 3.88	3 7.50	8 4.73
Total	129 100.00	40 100.00	169 100.00

Use `tab` (`tabulate`) to draw a 2×2 contingency table, with added column percentages (`col`) to read proportions of female leaders.

Stata implementation: prtest

```
. prtest leader, by(parliamentary)

Two-sample test of proportion                   0: Number of obs =      129
                                                1: Number of obs =       40

------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           0 |   .0387597   .0169946                      .0054509    .0720685
           1 |       .075   .0416458                     -.0066243    .1566243
-------------+----------------------------------------------------------------
        diff |  -.0362403   .0449799                     -.1243993    .0519187
             |  under Ho:   .0384317    -0.94   0.346
------------------------------------------------------------------------------
        diff = prop(0) - prop(1)                                  z =  -0.9430
    Ho: diff = 0

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(Z < z) = 0.1728         Pr(|Z| < |z|) = 0.3457          Pr(Z > z) = 0.8272
```

The difference in proportions is statistically insignificant at $p < .05$, even though it might be substantively significant (Type II Error).