Day 4: Research Design issues in textual studies

Kenneth Benoit

Essex Summer School 2012

July 12, 2012

General Issues

1. Validity: does a measurement reflect the truth of what is being measured?
2. Reliability: does repetition of a research procedure produce stable results?
3. Replicability: can a text analysis procedure be repeated at all?
4. Uncertainty: what is the variability of our estimates?
5. Precision: How exact are the estimates from our procedure?
6. Accuracy: How closely do our estimates correspond to the truth?

Tradeoff: Reliability *contra* validity

- Reliability refers to the dependability and replicability of the data generated by the text analysis method
- Validity is the quality of the data that leads us to accept it as "true," insofar as it measures what it is claimed to measure
- In text analysis, these two objectives frequently trade off with one another, since only human judgment can (ultimately) ensure validity, but human judgment is inherently unreliable
- Each concept has many variations, and in the case of reliability, several measures that can be applied
- Validity is the hardest to establish, since questions can always be raised about human judgment

Examples of tradeoffs

- Examples in coding text units:
  - Perfectly reliable procedure: Code all text units as pertaining to "Economic growth: positive"
  - Perfectly valid: Get a Nobel Prize laureate in economics to classify each text unit
- Examples in unitizing a text:
  - Perfectly reliable: Have a computer parse all texts into *n*-grams, such as words, pairs of adjacent words, etc. based on pre-defined rules (space is a delimiter, etc.)
  - Perfectly (?) valid: Have expertly trained humans parse the text into "quasi-sentences"

Validity: Definitions

Validity is a quality of a research measure or conclusion.

- The extent to which an empirical measure adequately reflects what humans agree on as the real meaning of a concept (Babbie 1995)
- The idea that the research should speach as truthfully as possible to as many as possible (Riffe, Lacy, and Fico 1998)
- Has many different varieties, which we will discuss below
- Challenge in text analysis is establishing, by various means, the validity of the procedure and its findings
- Invalid procedures are very easy to devise: for instance counting the number of words starting with "z"!

Validity: Types

"Face validity" The extent to which a measure seems plausibly to represent a concept, "on the face of things". Implies a result on which inter-subjective agreement would not be difficult to obtain.

Criterion validity The extent to which a measure taps an established standard or important behavior that is external to the measure. Assumes that standards both exist and can be agreed upon. Has additional nuances such as "concurrent" and "predictive" versions.

Content validity The extent to which a measure reflects the full domain of the concept being measured. Example: the coverage of 26 ingredients in the CMP's left-right policy scale.

Validity: Types

"Social" validity (Krippendorff) Acceptance criterion based on contribution to important public issues, social relevance, etc.

Construct validity The extent to which a measure is related to other measures in a way consistent with hypotheses derived from theory. Established through cross-validation with other, independent measures. In content analysis of political policy, for instance, this frequently means comparing results with survey or expert survey measures.

External validity The generalizability of the findings, in terms of whether they also hold true for other settings, times, etc. Established through the representativeness of the sample, and through extension to additional contexts through further research.

Typology of validation efforts in content analysis

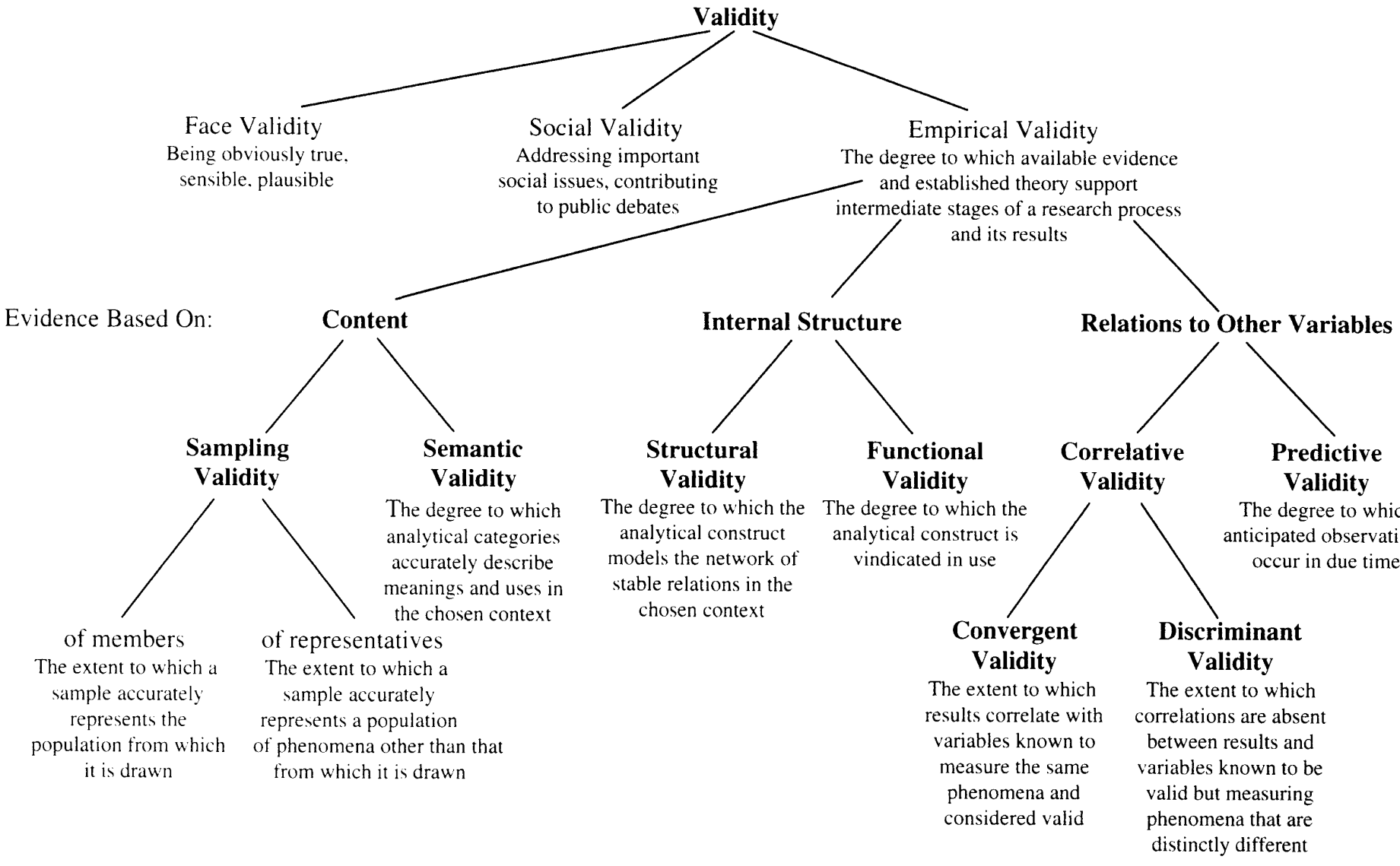


Figure 13.1 of Krippendorff)

Reliability: Definitions

Reliability in essence means getting the same answers each time an identical research procedure is conducted.

- The extent to which a research procedure yields the same results on repeated trials (Carmines and Zeller 1979)
- The assurance that data are obtained independently of the measuring event, instrument, or person, and that remain constant despite variations in the measuring process (Kaplan and Goldsen 1965)
- Interpretivist conception: Degree to which members of a designated community agree on the readings, interpretations, responses to, or uses of given texts or data (Krippendorff)

Importance of Reliability

- In text analysis (and most other forms of empirical analysis), unreliable procedures yield results which are meaningless.
- Typically measures in terms of agreement between two human coders, when referring to hand-coded content analysis
- Computerized methods have largely removed this concern, inasmuch as they are mechanical procedures that yield the same results each time the procedure is repeated.

Types of reliability

Distinguished by the way the reliability data is obtained.

Type	**Test Design**	**Causes of Disagreements**	Strength
Stability	test-retest	intraobserver inconsistencies	weakest
Reproducibility	test-test	intraobserver inconsistencies + interobserver disagreements	medium
Accuracy	test-standard	intraobserver inconsistencies + interobserver disagreements + deviations from a standard	strongest

Reliability test designs

Test-retest The same text is reanalyzed/reread/reclassified, or the same measurement is repeatedly applied to the same set of texts. Goal is to establish inconsistencies. (Establishes *stability*)

Test-test Two or more individuals, working independently, apply the same analysis instructions to the same texts, to compare intraobserver differences. (Establishes *reproducibility*).

Test-standard The perfomance or one or more procedures is compared to a procedure that is taken to be correct. Deviations from a ("gold") standard are then recorded. (Establshes *accuracy*.) Typically used in coder training, or training of automated (computer-based) procedures.

Designing reliability checks in practice

- Repeating the procedure on the sample data
- Using independent tests from separate coders
- Can a "gold standard" be identified?
- Split-design tests
- Example: CMP
  - Same coders repeat own codings
  - Different coders code same test
  - The "reliability" coefficient reported in the dataset is correlation of category percentages obtained by a coder on the training document used by CMP versus the master "gold standard" version of the coding done by Andrea Volkens

Measures of agreement

- Percent agreement Very simple: (number of agreeing ratings) / (total ratings) * 100%
- Correlation
  - (usually) Pearson's r, aka product-moment correlation
  - Formula: $r_{AB} = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{A_i - \bar{A}}{s_A} \right) \left(\frac{B_i - \bar{B}}{s_B} \right)$
  - May also be ordinal, such as Spearman's rho or Kendall's tau-b
  - Range is [0,1]
- Agreement measures
  - Take into account not only observed agreement, but also *agreement that would have occured by chance*
  - Cohen's κ is most common
  - Krippendorf's α is a generalization of Cohen's κ
  - Both range from [0,1]

Reliability data matrixes

Example here used binary data (from Krippendorff)

Article:	1	2	3	4	5	6	7	8	9	10
Coder A	1	1	0	0	0	0	0	0	0	0
Coder B	0	1	1	0	0	1	0	1	0	0

- A and B agree on 60% of the articles: 60% agreement
- Correlation is (approximately) 0.10
- Observed *dis*agreement: 4
- Expected *dis*agreement (by chance): 4.4211
- Krippendorff's $\alpha = 1 - \frac{D_o}{D_e} = 1 - \frac{4}{4.4211} = 0.095$
- Cohen's κ (nearly) identical

Reliability and validity differences

- Reliability can be established through tests as a part of a research procedure; validity cannot be established through the same sort of (repetition) tests.
- Validity concerns substantive *truths*, whereas reliability is mainly procedural.
- Unreliability limits the chance of obtaining valid results, in the sense that procedures whose results cannot be trusted are less likely to be true.
- Reliability is no guarantee of validity, since reliable procedures can be consistently wrong, even when these procedures involve human judgment.

Additional (related) concepts

Generalizability The extent to which findings may be applied to cases other than those from which the research is immediately taken, for instance from a sample to a population. (We will subsume this under "external validity".)

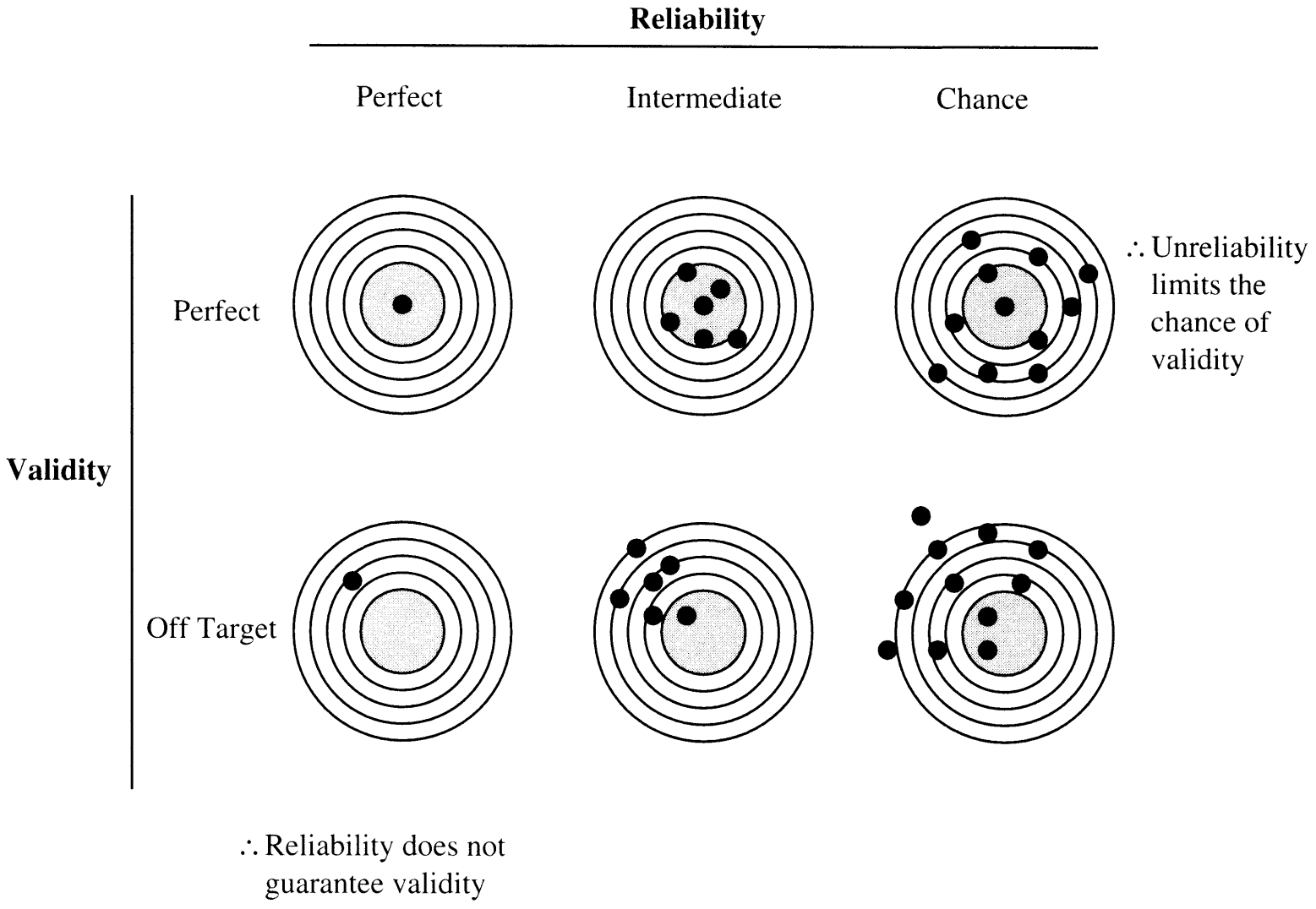
Precision The fineness of distinction or level of measurement. For instance, measuring time in morning/afternoon versus HH:MM:SS.

Accuracy The extent to which a measurement corresponds to the truth – usually determined by whether it is free from bias, but also affected by reliability.

These last two concepts also trade off with one another: highly precise measures are less likely to be accurate.

Interrelation of additional concepts

(From Krippendorff Figure 11.1)



The design of the experiment

- Data: 14 speeches from the debate on Irelands 2010 budget (FF+Greens vs FG+Lab+SF)
- Subjects: 18 human readers, mostly PhD students (LSE and TCD)
- Task: Identify speaker positions, directly and by pairwise comparison and indicate uncertainty
- Questions: Does the model recover human positioning? What is appropriate certainty?

Walk through the paper...

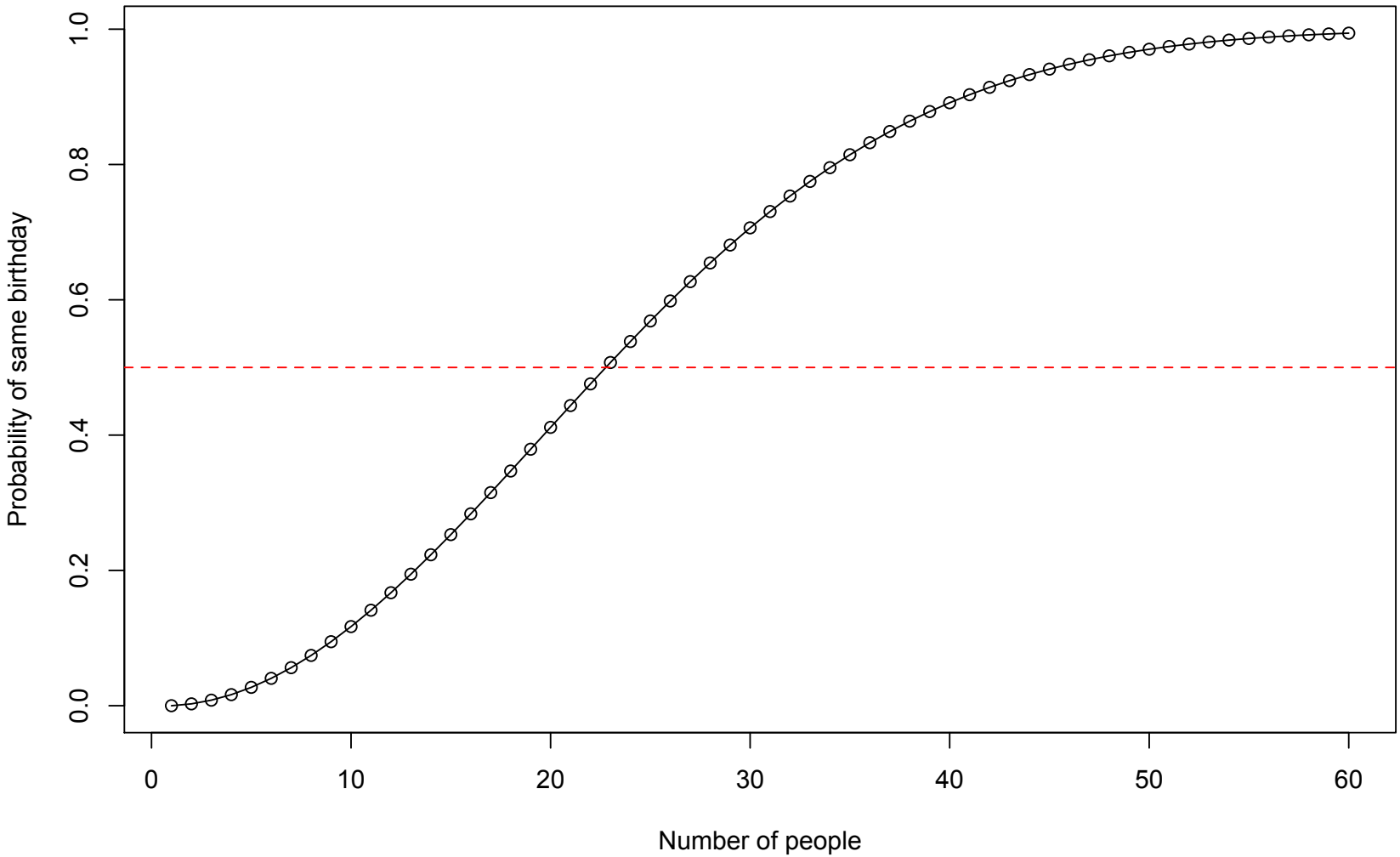
Another probability puzzle: The Birthday Problem

The Birthday Problem: What is the probability that two people in this room will have the same birthday?
One of the most famous problems in combinatorics and probability.
What is the probability that in a room of n people, any two have the same birthday?

- We start with (wrong!) assumptions: no leap years, no twins, no seasonal or weekday variations, all birthdates equally likely
- Rephrase question: What is probability that no two of n people will share a birthday?

The Birthday Problem

- Probability is 0 with 366 people
- Probability is 1.0 with 1 person, or $\frac{365}{365} = 1.0$
- Probability for two people is: $\frac{365}{365} \cdot \frac{364}{365} = 0.9973$
- Probability for three people is: $\frac{365}{365} \cdot \frac{364}{365} \cdot \frac{363}{365} = 0.9918$
- Formula for n people is:
  $\frac{365-1}{365} \cdot \frac{365-2}{365} \cdot \ldots \cdot \frac{365-1-n}{365}$
- alternatively
  $\left(1 - \frac{365}{n}\right) \cdot \frac{1}{365^n} = \frac{365!}{(365-n)!265^n}$
- Crosses 0.50 at just 23 people!
- More than 0.75 at 30 people, and 0.99 at 57 people


Probability of same birthday
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
60
Number of people

Working in R: Birthday problem example

- Formula: $1 - \frac{365!}{(365-n)!365^n}$
- In R, w can use the `factorial()` function
- So for $n = 10$:
  `1 - (factorial(365) / (factorial(365-n) * 365^n))`
- Does this work? No – numbers too big!
- How to solve this: use logarithms and `lfactorial()`:
  `1-exp(lfactorial(365) - lfactorial(365-n) - n*log(365))`

Working in R: Birthday problem example code

```
lbdp <- function(n) {
  1 - exp(lfactorial(365) - lfactorial(365-n) - n*log(365))
}

x <- 1:60
plot(x,lbdp(x))

plot(x, lbdp(x), type="o",
     xlab="Number of people",ylab="Probability of same birthday")
abline(h=.5, lty="dashed", col="red")
```