Statistiek I

Proportions aka 'Sign Tests'

John Nerbonne

CLCG, Rijks*universiteit* Groningen

http://www.let.rug.nl/nerbonne/teach/Statistiek-I/

Proportions aka 'Sign Test'

The relative frequency of nominal (categorical) data is sometimes an issue.

Example: percentage of women customers for an e-business site. Sex is male vs. female, therefore categorical data.

Question: is the percentage of women significantly greater at one sort of site (e.g., films) as opposed to another (e.g., music)?

Two Approaches

1. M&M (5.1): Proportions (percentages) are **not** numerical data.
   t-tests do **not** apply.
   We **can** analyze proportional data using the **binomial distribution**, from which a *z*-**value** can be derived.
2. Proportions may be viewed as numerical data. Use *t*-test.
   See `http://home.clara.net/sisa/`

Numbers and Fractions

$B(n, p)$ BINOMIAL DISTRIBUTION of n events, all with p chance

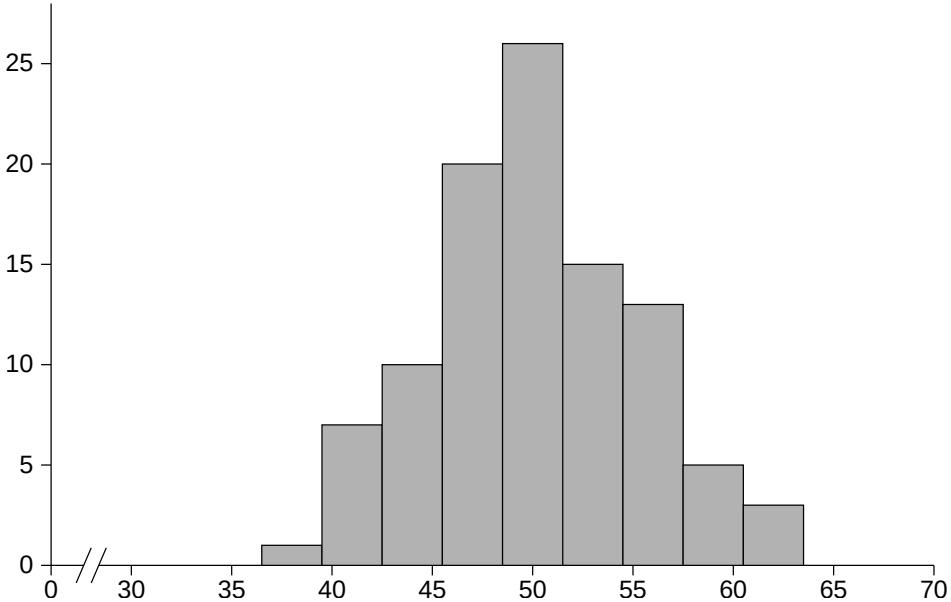
- fixed number of observations n
- each observation can classified in one of *two* ways success & failure
- all independent: chance of success p same throughout

Experiment: Repeat 100 times:
— toss a coin 100 times, record number of heads $B(100, 0.5)$

Also known as BERNOULLI TRIALS
Not discussed in Field!

Numbers and Fractions

Experiment: Repeat 100 times: record number of heads in 100 tosses



In large samples, the binomial distribution resembles the normal (z usable).

Binomial Chances

n repetitions of independent events with chance p will have a binomial distribution $B(n, p)$

- each sequence including k successes will include $n - k$ failures and will have the probability

$$\underbrace{p \cdots p}_{k} \cdot \underbrace{(1-p) \cdots (1-p)}_{n-k} = p^k \cdot (1-p)^{n-k}$$

- there are $\binom{n}{k}$ sequences w. k successes, $n - k$ failures

the chance of k successes is

$$P(N = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

Binomial Distributions

n repetitions of event with prob p show dist. $B(n, p)$

Example: you write software for information retrieval. The prob. of a randomly returned title (from a specialist library) being relevant to a query is $p = 0.0003$ Your software never returns more than 100 titles, so the prob. of there accidentally being a relevant title in a query-response is 0.03

If your software is used in an experiment with 100 queries, what is the chance that *no* query-response will contain any relevant title?

$$\begin{aligned} P(N = 0) &= \binom{100}{0}(0.03)^0(0.97)^{100} \\ &= 1 \cdot 1 \cdot (0.97)^{100} \\ &= 0.048 \end{aligned}$$

Binomial Distributions

Example: you are investigating *anomia*, a form of aphasia in which naming skills are lost. It's incidence in the population is $p = 0.0015$. What is the prob. of there being 10 in a random sample of 500?

$$P(N = 10) = \binom{500}{10}(0.0015)^{10}(0.9985)^{490}$$

Binomial Expectation, Standard Error

In Bernoulli trials with n repetitions, each with chance p.

Let $\hat{p}$ be the proportion of successes seen in the sample.

Expectation: $\hat{p} = p$

Standard Error: $\text{SE}_p = \sqrt{p(1-p)/n}$

Use SE_p to find the z-value of a proportion with respect to a hypothesized p.

$$z = \frac{\hat{p} - p}{\text{SE}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}}$$

Example

Example: As before, women customers for an e-business site, again testing at the 0.05-level whether 20% of customers are women.

$H_0 : p = 0.2$ (20% of customers are women)
$H_a : p \neq 0.2$ (it isn't true that 20% of the customers are women)

Sample: 100 customers selected at random. We therefore derive a SE for sample size 100:

$$\begin{aligned} \text{SE}_p &= \sqrt{p(1-p)/n} \\ &= \sqrt{0.2(0.8)/100} \\ &= \sqrt{(0.16)/100} \\ &= \sqrt{0.0016} \\ &= 0.04 \end{aligned}$$

Example Results

Results: 12 of 100 customers selected at random are women, i.e., $\hat{p} = 0.12$. We use this to test the hypothesis that 20% of the customers in the population are women ($p = 0.2$). Using $\text{SE}_p = 0.04$, we can derive a z for sample size 100:

$$z = \frac{\hat{p} - p}{\text{SE}} = \frac{\hat{p} - p}{\sqrt{p(1-p)/n}} = \frac{0.12 - 0.20}{\sqrt{0.2(0.8)/100}} = \frac{-0.08}{0.04} = -2$$

Since we know that there is ≤ 0.05 chance that z is this extremely different from 0, this result should expected less than 5% of the time, if H_0 were true. $p \leq 0.05$.

Conclusion: Reject H_0. It is not true that 20% of the customer of the e-business site are women. The result is significant at the $p = 0.05$-level.

Reasoning about Proportions

There should be minimally ten examples of success and ten examples of failure for this procedure to be applied, i.e., $p \cdot n \geq 10$ and $(1 - p) \cdot n \geq 10$.

This can mean that large samples need to be examined in case p is very large or very small.

Proportions in Two Samples

You can also test hypotheses about proportions in two samples. In this case you use both standard proportions and both sample sizes to determine the standard deviation:

$$SE_p = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

Example: Suppose you wished to compare the proportion of women customers at one e-business site as opposed to another, asking the question as to whether the proportions are significantly different.

$H_0 : p_1 = p_2$ (proportions about the same at both sites)
$H_a : p_1 \neq p_2$ (proportions different)

Two Samples — Results

You collect data from two sites, determining that 12 of the 100 customers from the sample at one site were women and 22 of 100 at another. To test the hypothesis that these are different, you derive a z-value based on the null hypothesis (that they're the same).

$$z = \frac{p_1 - p_2}{\mathrm{SE}_p} = \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

Two Sample — Calculations

$$
\begin{aligned}
z = \frac{p_1 - p_2}{\text{SE}_p} &= \frac{p_1 - p_2}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \\
&= \frac{0.12 - 0.22}{\sqrt{\frac{0.12(0.88)}{100} + \frac{0.22(0.78)}{100}}} \\
&= \frac{-0.1}{\sqrt{\frac{0.1056}{100} + \frac{0.1716}{100}}} \\
&= \frac{-0.1}{\sqrt{0.001056 + 0.001716}} \\
&= \frac{-0.1}{\sqrt{0.002772}} \\
&\approx \frac{-0.1}{0.05265} \approx -1.9
\end{aligned}
$$

$P(|z| \geq 1.9) = 0.0574 \ (\not\leq 0.05)$

Conclusion: There is insufficient evidence to reject H_0 at $p \leq 0.05$-level. The underlying population proportions may be same.

Binomial Analysis

- Use binomial analysis for proportions of nominal data (no t-test!)
- Let p be the chance of success, $(1 - p)$ chance of failure.
- Both $p \cdot n \geq 10$ and $(1 - p)n \geq 10$
- Use binomial SE

$$\mathrm{SE}_p = \sqrt{p(1-p)/n} \quad \text{(single sample)}$$

$$\mathrm{SE}_p = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad \text{(two samples)}$$

Example: Disambiguation Accuracy

Tanja Gaustad (2004) developed software to determine intended word senses in text. She tested her software on 55,000 examples of words for which the disambiguation was known.

Note that diferent methods were applied to the same data (paired data). This is like the McNemar test (see labs).

When she contrasted methods, there were three possible outcomes:

1. The methods agreed (these cases were discounted).
2. The first method was right, and the second wrong.
3. The second method was right, and the first wrong.

In a typical contrast the methods agreed in almost all cases (all but 500). She applied a binomial analysis to these cases to inspect for significance.

Paired Sign Test

This is also known as the PAIRED SIGN TEST because the methods are contrasted on exactly the same material.

We reason from a background assumption that the methods differ on the basis of chance. In that case the first will be right in 50% of the cases in which they differ.

How many more examples does the better method need in order to be significantly better?

$$\sigma = \sqrt{np(1-p)} = \sqrt{500 \cdot 0.5 \cdot 0.5} = \sqrt{125} \approx 11.2$$

We ask how many more than $n \cdot p = 250$ did a method need to analyze correctly in order to be significantly better ($\alpha = 0.05$)

Paired Sign Test

Our H_0 is that the two methods differ at a chance level, the alternative that one is better than the other.

$$\begin{aligned} P(z \geq 1.65) &= 0.05 \\ P(c \geq \mu + 1.65\sigma) &= 0.05 \\ P(c \geq 250 + 1.65(11.2)) &= 0.05 \\ P(c \geq 268.5) &= 0.05 \end{aligned}$$

A method that only got about 40 more examples right than a competitor is thus better to a statistically significant degree.

This corresponds to only 0.07% improvement in accuracy of processing (40/55,000).

Technical note: The McNemar test is essentially a two-cell, one-df χ^2 comparison, *not* based on the binomial. No need for Yate's continuity correction with binomial.

t-Test on Proportions

If we are willing to regard counts as numerical scores, then we have a reason to use the *t*-test.
We illustrate this technique on the question of whether there is a difference in the proportion (chance) of inflection errors made by Frisian children depending on whether they grow up in an exclusively Frisian setting or a mixed Frisian-Dutch setting.

This is a **two-sample** application.

t-test for Proportions in Two Samples

Nynke van den Bergh studies children acquiring Frisian. There are two groups:

- children who hear only Frisian at home and in child-care settings
- children who hear Frisian at home and Dutch in child-care settings.

The question is whether the mixed setting will lead to more interference errors—errors whether the child uses a Dutch pattern instead of a Frisian one.

Van den Bergh has studied patterns of the type:

Produced	**Target**	**Translation**
Gean mei boartsie	gean mei boartsjen	'go play along'
Ik kin swimmen	ik kin swimme	'I can swim'

Frisian/Dutch Interference Hypothesis

Van den Bergh's null hypothesis is of course, that there is **no** difference in the proportion of incorrect inflections in the two populations (of expressions of inflection among children from the purely Frisian environment on the one hand as opposed to the children from the mixed environment on the other). Her alternative hypothesis is that the children from the mixed environment show more errors due to interference.

The hypothesis is therefore **one-sided**:

H_0 $p_F = p_M$, where p_F is the error percentage of children in the purely Frisian environment, p_M the error percentage of children in mixed environment.

H_a $p_F < p_M$

Frisian/Dutch Interference Data

Van den Bergh's data for kids 5 years, 11 months old:

Setting	Correct	Incorrect
Pure Frisian	85 (97.7%)	2 (2.3%)
Mixed	167 (89.8%)	19 (10.2%)

We wish to assume that there is **no** difference in the proportions in the two populations (the population of kids' expressions in the pure setting and their population in the mixed setting), and ask how likely these samples are given that assumption.

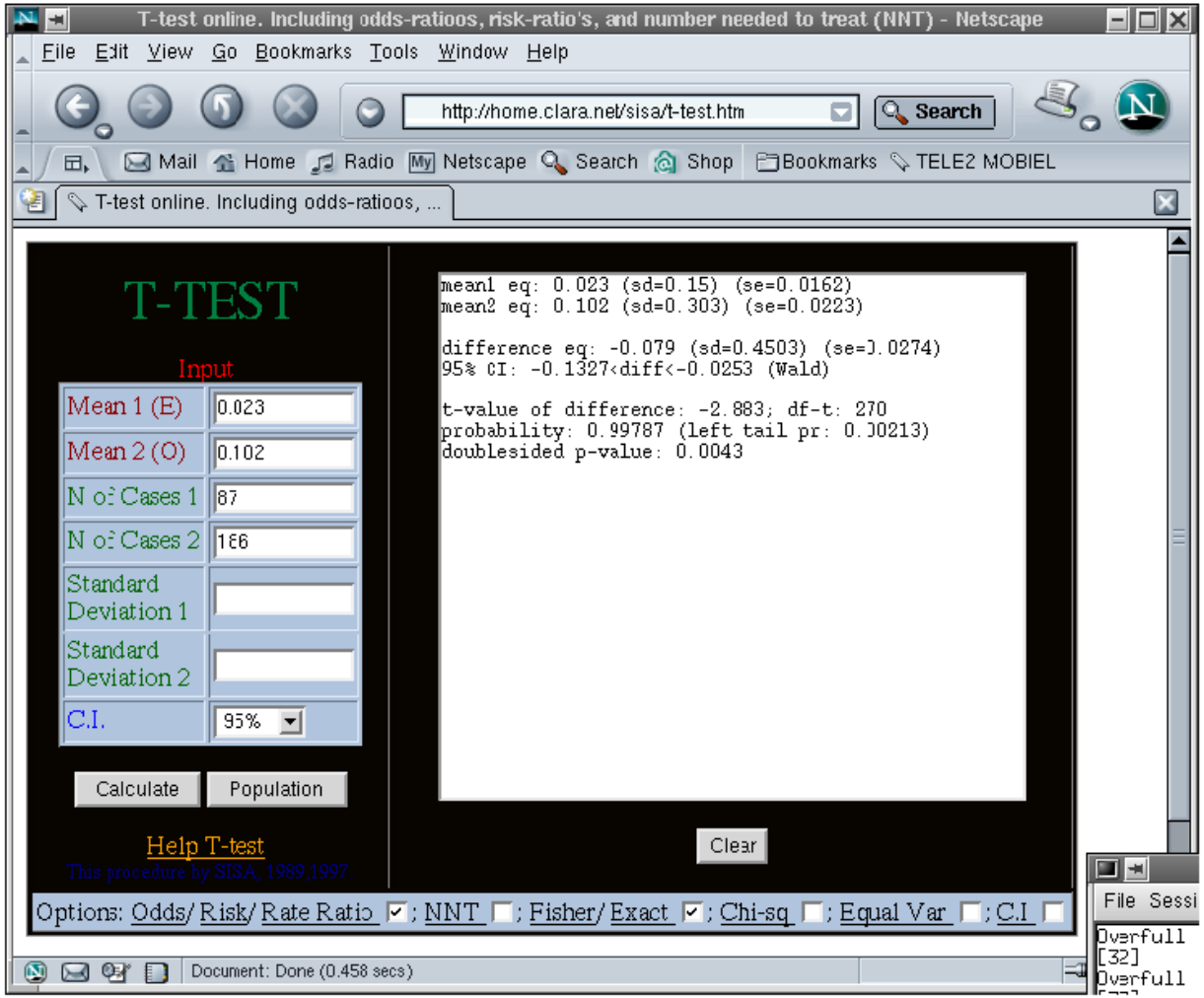
Frisian/Dutch Interference

We can test whether two proportions are significantly different at several online web-sites for statistics, e.g.

`http://home.clara.net/sisa/t-test.htm` Sisa

We only need to input the proportions, 0.023 error rate vs. 0.102 error, and the total number of elements, 87 and 186, respectively.

Sisa Calculations



Sisa Results

For this period, the Sisa web site calculated the following:

```
mean1 eq: 0.023 (sd=0.15) (se=0.0162)
mean2 eq: 0.102 (sd=0.303) (se=0.0223)
...
t-value of difference: -2.883; df-t: 270
probability: 0.99787 (left tail pr: 0.00213)
doublesided p-value: 0.0043
```

We're interested in the one-sided p value: whether the children in purely Frisian environments make **fewer** mistakes than those in mixed environments.

Indeed they make significantly fewer mistakes at this period ($p < 0.01$).

Frisian/Dutch Interference—Reasoning

We are asking about the chance of two samples from the sample population differing. To estimate this, we need an estimate of the standard deviation and the standard error.

We use the estimates of sample standard error from the binomial distribution.

$$
\begin{aligned}
SE_p &= \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \quad \text{(two samples)} \\
SE_p &= \sqrt{\frac{0.023(0.977)}{87} + \frac{0.102(0.898)}{186}} \\
SE_p &= \sqrt{\frac{0.022}{87} + \frac{0.091}{186}} \\
SE_p &= \sqrt{0.00025 + 0.0005} \\
SE_p &= 0.0274
\end{aligned}
$$

The *t* value is based on this.

Two Samples Proportions — t-value

We can now show

$$\begin{aligned} t = \frac{p_1 - p_2}{\mathrm{SE}_p} &= \frac{0.102 - 0.023}{0.0274} \quad \text{(see last slide)} \\ &= \frac{0.079}{0.0274} \\ &= 2.88 \end{aligned}$$

If we compare this to the t-tables (M&M, Table E, p.705) for 80 Deg. Freedom (smaller of n_1, n_2), we derive $p = 0.0025$.

The Sisa site is less conservative in estimate of deg. freedom (using $n_1 + n_2$), and arrive at $p = 0.0021$.

Two Samples Proportions

The data is usually nominal (categorical). **Example:** percentage of women customers for e-business sites. Sex is male vs. female, therefore categorical data. Question in that situation might be: does a significantly greater percentage of women visit one site as opposed to another.

Two Approaches

1. M&M (5.1) recommend using binomial distribution to analyse proportions, estimating a z value from that.
2. Others view proportions as numerical data, recommending t-test. See `http://home.clara.net/sisa/`

Next

Review for Exam!