Statistics for EMCL
week 4

Tamás Biró

Humanities Computing

University of Groningen

t.s.biro@rug.nl

This week

- Inference for proportions (M&M ch. 8)
- Includes an intro to probability theory

Summary so far

- z-test: if population σ known.

- t-test:

  - One-sample t-test, a.k.a. single sample t-test.
  - Matched pairs t-test, a.k.a. paired data t-test.
  - Two-sample t-test, a.k.a. independent samples t-test.

Same for confidence intervals.

A note for "senior supervisors"

- Two-sample t-test as we have learned a.k.a. **Welch-test** (SPSS "equal variances not assumed"): nowadays easy to perform using software and highly preferred by M&M.

- Traditional approach (not recommended by M&M, not required for the exam, but be aware of it):
  - First, **F-test** for equality of spread (M&M 7.3)
  - Then, **pooled two-sample t-test** (M&M 7.2) (SPSS "equal variances assumed")

Summary so far

Statistic's encountered:

- Sample size, min, max, range, mode, median, Q1, Q3, IQR, percentiles, mean, standard deviation ($n-1$ and n), standard error, z-statistic, t-statistic (for one and two samples).

Summary so far

One	case	sample
Measure	value of variable X	statistic (e.g. mean)
Many	distribution of value in a sample	distribution of stat. in a sample of samples
All	distribution of value in population	sampling distribution of the statistic

A useful distinction: type vs. token

How many letters are there in an English text?

- Token: there are 42243 letters.
- Type: there are 26 different letters.

Several copies of the same type are different tokens.

Histogram: show how many tokens belong to the same type.

Even more levels

Compare mean-utterance-length across languages.

- A case = a language.
- Variable = mean-utterance-length.
- A value = mean of the length of different utterances in language L.
- Statistic = mean of MLU in a sample of n languages.
- Mean of the sampling distribution of mean = mean of sample means of MLU.

And now: proportions

Proportion

- Variable is "Boolean": yes/no (success/failure, true/false, correct/incorrect).
- Population's parameter: proportion p
  - $p\%$ is "success",
  - $1 - p$ (that is, $100 - p\%$) is "failure".

Probability and dices

Don't take notes!

Example: throwing a dice

I threw a dice 60 times and only 8 times did I get a "6".

- Can I conclude with significance level of 0.05 that the dice is biased?

First approach to probability

If I throw an unbiased dice "infinite" times, I get "1" in $1/6$ of the cases, "2" in $1/6$ of the cases, etc.

- Population: many-many experiments/observations.
- Parameter p: ratio of "6"'s in the population.
- Probability of throwing a "6" is p.

Law of large numbers

(M&M 4.4, p. 274)

- An **observation**: several possible outcomes, called **events**.
- Event has probability associated with it.
- Repeat the observation many times:
- the relative frequency of each event will converge to its probability.

Towards a binomial distribution

- Observation: throw a dice.
- Event 1: "6" (probability $p = \frac{1}{6}$).
- Event 2: "not 6" (probability $q = \frac{5}{6}$).
- Event 1 and event 2: mutually self-exclusive, and not other possibility; therefore, $p + q = 1$.

Population

- Population: as many observations as you want.

- Law of large numbers: then, ratio of event "6" will be as close to $p = \frac{1}{6}$ as you wish. Similarly, ratio of even "not 6" as close to $q = \frac{5}{6}$ as you wish.

Sample

- Sample: repeating the observation "only" n times.

- Question: what is the probability of having k times "6" and $n-k$ times "not 6"?

Meaning of probability again

- Meaning of this latter probability: if you collect many samples of n observations each, how many of these samples will contain “6” k times?

- Probability of event: its frequency in population of observations; probability of sample: its frequency in population of samples.

NB: Exact mathematical approach sees frequency as the consequence and not as the definition of probability.

Probability of sample

Sample of size n: what is the probability of having k times "6" and $n-k$ times "not 6"?

- Probability of getting k times "6" in a row:
  $p \cdot p \dots \cdot p = p^k = \left(\frac{1}{6}\right)^k$.

- Probability of getting $n-k$ times "not 6" in a row:
  $q \cdot q \dots \cdot q = q^{n-k} = \left(\frac{5}{6}\right)^{n-k}$.

(Cf. multiplication rule for independent events, M&M 4.2.)

Probability of sample

In a sample of size n, what is the probability of having k times "6" and $n-k$ times "not 6"?

- Probability of getting k times "6" and then $n-k$ times "not 6": $p^k \cdot q^{n-k}$.
- Different orders of "6" events and "not 6" events.

Binomial coefficient

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

Read: "binomial coefficient n choose k".
Meaning: the number of ways k elements can be chosen from n elements.
See Table C of M&M.

NB: Factorial: $x! = 1 \cdot 2 \cdot ... \cdot (n-1) \cdot n$.

Binomial distribution

(Cf. M&M 5.1.)

- **Binomial distribution** $B(n,p)$: n observations, with probability p of success in each observation.

- **Binomial probability**: probability of having exactly k observations of success:
  $P(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$

What concerns us:

- Mean of binomial distribution: $\mu = np$.
- Variance of binomial distribution: $\sigma^2 = np(1 - p)$.

Back to the (un)biased dice

Take notes again!

Example: throwing a dice

I three a dice 60 times and only 8 times did I get a “6”.

- Can I conclude with significance level of 0.05 that the dice is biased?

Statistical test

- Null hypothesis: dice is unbiased:
  H_0: $p = \frac{1}{6}$.

- P-value: the chance of 8 or less times "6" out of 60 observations:

$$P = P(X = 0) + P(X = 1) + ... + P(X = 8)$$

for binomial distribution $B(n = 60, p = \frac{1}{6})$.

In practice

- Easy in theory, more difficult in practice.
- Software will do it for you for small n.
- Medium high n: *plus-four* estimate

  (not required, but know where to find it in M&M).
- High n: approximate with Normal distribution.

Good news:

Normal distributions

Inference for proportion

- Population: p of it is "success" and $1 - p$ is "failure".
  Example: $p = \frac{1}{6}$ is "6" for dice.

- Sample of size n: "success" X times.

- **Sample proportion**: $\hat{p} = \frac{X}{n}$.

- Question: can $\hat{p}$ approximate p?

Use z-procedures

- Unknown population proportion: p
- Sampling distribution: (approx.) Normal.
- Mean of sampling distribution: np.
- Variance of sampling distribution: $\sigma^2 = np(1-p)$.
- Use $\hat{p}$ for p if needed.

Large-sample Confidence interval for population proportion

- Sample proportion: $\hat{p} = \frac{X}{n}$.
- Standard error: $\mathrm{SE} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
- Find z^* for C.
- Confidence interval: $\hat{p} \pm z^*\mathrm{SE}$.

Choosing a sample size

See M&M.

- Either guess an approximation for p (based on common sense or past experience; if not exact, not much can be lost),
- or take $p = 0.5$, which is the worst case (if p close to 0 or to 1, you get a sample size much larger than needed).

Large-sample significance test for population proportion

- Null hypothesis: H_0: $p = p_0$.
- Use p_0 in standard error.
- z-statistic: $z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$
- Find P-values, one-sided or two-sided.

Two populations

- Population 1: parameter is proportion p_1.
- Population 2: parameter is proportion p_2.
- Sample 1 of size n_1 randomly drawn from population 1.
- Sample 2 of size n_2 randomly drawn from population 2.

- Count of success in sample 1: X_1,
sample proportion in sample 1: $\hat{p}_1 = \frac{X_1}{n_1}$.

- Count of success in sample 2: X_2,
sample proportion in sample 2: $\hat{p}_2 = \frac{X_2}{n_2}$.

Question: relation of p_1 to p_2, estimated from $\hat{p}_1$ and $\hat{p}_2$.

Frisian/Dutch interference

Nynke van Bergh studies children (5;11 year-old) acquiring Frisian. There are two groups of children:

- Frisian at home and Frisian in child-care.
- Frisian at home and Dutch in child-care.

Rate of use of Dutch patterns instead of Frisian ones.

Frisian/Dutch interference

Q: more interferences in bilingual environment?

- Population 1: sentences by children in group 1
- Population 2: sentences by children in group 2
- Proportions of correct and incorrect sentences:

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

Frisian/Dutch interference

One-sided hypothesis, because of theory (and not after having seen the data):

- Null hypothesis: H_0: $p_F = p_M$
- Alternative hypothesis: H_a: $p_F < p_M$

Frisian/Dutch interference

- http://home.clara.net/sisa/t-test.html
- Proportions 0.023, 0.102; total number of elements 87 and 186.

Significance level: $P < 0.01$.

Statistical procedures for two proportions

- Large-sample significance test: is $p_1 = p_2$?
- Large-sample confidence interval: how much is $p_1 - p_2$?
- (Relative risk: how much is $\frac{p_1}{p_2}$?)

See details M&M ch. 8.2, use SPSS.

Next week:

- Two variables, chi-square test (M&M 2 and 9.1)
- Use reading week to catch up with reading
- Read also M&M chapter 2 (focusing on concepts, and ignoring maths).