Statistics for EMCL week 3

Tamás Biró

Humanities Computing

University of Groningen

`t.s.biro@rug.nl`

Technicalities

- Sorry for delay in correcting assignments.
- Feedback to assignment 1 in email + on web.
- No lab on Thursday, October 9, no lecture on Wednesday, October 15.
- Next week: lecture Thursday, October 16, 11.15, room 12.0120. Then lab, as usual.

Technicalities

- Everything in Moore & McCabe as described on website (weekly "readings"). Except: sections with a *, and certain mathematical details (if emphasized during lecture). For those using older editions of M&M: some details are missing (e.g., section 3.4) and some are located elsewhere.

- Questions? Extra Q&A session after break?

This week

- Confidence intervals (z and t).
- Statistical tests (z and t).

Procedure of inference

- Sample $\longrightarrow$ statistic $\longrightarrow$ parameter estimated.
- Trustworthiness of procedure: what would happen if we repeat this procedure many times?
- **Sampling distribution of a statistic**: distribution of the statistic, if taken from all possible samples.

Example: task 2 in lab 2.

Inferential statistics

$$
\begin{array}{ccc}
\mathit{population} & \longrightarrow & \text{sample1, sample2, sample3...} \\
\downarrow & & \downarrow \\
\mathit{parameter} & \longleftarrow & \text{sampling distribution of statistic}
\end{array}
$$

But usually money for one sample only!

We focus on

- Distribution can be any (better if Normal).
- Samples: SRS (random, etc.).
- Parameter: mean μ.
- Statistic: mean $\bar{x}$.

Central Limit Theorem

- Given population with any distribution.
- Its mean is μ. Its standard deviation is σ.
- Draw a *simple random sample* (SRS) of size n.
- Calculate sample mean $\bar{x}$. Then **CLT**:
- Sample distribution of $\bar{x}$ is Normal: $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- This theorem is approximately true if original population is not Normal, but n is large.

Central Limit Theorem

- Even if we do not know the distribution of some variable in the entire population,
- we know how the empirical mean $\bar{x}$ of any large random sample behaves:
- it is distributed around the mean μ of the population,
- and it follows a Normal distribution of mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$.

How to use this knowledge?

Statistical procedures for inference

- **Confidence interval:** unknown mean μ of population is in the interval $\bar{x} \pm m$ at a **confidence level** C.

- **Significance test:** the probability of the **null hypothesis** is "only" P, so **alternative hypothesis** is most probably true.

Statistical procedures for inference

- Requirements of a test. E.g., how much important is to have Normal distribution? approximately Normal distributions? (hence, use of Normal quantile plots). Often: less sensible to deviation from Normality if n is large.
- M&M 6.3: use and abuse of tests.

Statistic 1: mean $\bar{x}$

- Given sample of size n (= number of cases), and data points (values being measured on each case of the sample) x_1, x_2,... x_n,

- calculate *mean*: $\bar{x} = \frac{x_1 + \ldots + x_n}{n}$.

- Central Limit Theorem: $\bar{x}$ has sampling distribution $N(\mu, \sigma)$, where μ and σ are mean and stand. deviation of entire population respectively.

Statistic 2: z-statistic

- Given sample of size n and data points (values being measured on each case of the sample) x_1, x_2,... x_n,
- calculate z-*statistic*: $z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$.
- You can calculate z-statistic for any values μ, σ.
- Central Limit Theorem: if μ and σ are the true mean and st. dev. of entire population, then z has sampling distribution $N(0,1)$.

Statistic 3: t-statistic

- Given sample of size n and data points (values being measured on each case of the sample) x_1, x_2,... x_n,
- calculate t-$statistic$: $t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$.
- You can calculate this statistic with any value of μ, but s is sample standard deviation of sample!
- Maths: if μ is the true mean of entire population, then t has sampling distribution $t(n-1)$.

Two further statistics

- **Standard error** of sample: $\mathrm{SE}_{\bar{x}} = \frac{s}{\sqrt{n}}$.

  NB: see M&M for variations in terminology.

- **Degree of freedom** of sample: $df = n-1$

  Namely, nth data is not "free" once mean is fixed.

Further notes

- z if population σ known (M&M chapter 6): usually not realistic assumption.

- t if σ unknown (usually the case): approximation using s; but therefore $t(df)$, and not $N(0,1)$.

- $t(df)$: symmetric, bell-shaped curves, close to $N(0,1)$ (see M&M 7.1, p. 420).

Using Table D

- Last line: for Normal distribution $N(0, 1)$ (inverting Table A, please check it yourself!).
- Rest: for t-distribution $t(df)$.
- What is t^*/z^* such that
  - Area (probability) right of t^*/z^* is p?
  - Area between $-t^*/-z^*$ and t^*/z^* is C?

And now
derive
statistical procedures!

Confidence intervals for z-statistics

- Set **confidence level** C.
- Find corresponding z^*.
- $C\%$ of samples produce a z-statistics that is closer to 0 than z^*: $-z^* \leq z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \leq z^*$
- Hence: $\bar{x} - z^* \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z^* \frac{\sigma}{\sqrt{n}}$.
- In short: $\mu = \bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$.

Confidence intervals for t-statistics

- Set **confidence level** C.
- Find corresponding t^*.
- $C\%$ of samples produce a t-statistics that is closer to 0 than t^*: $-t^* \leq t = \frac{\bar{x}-\mu}{s/\sqrt{n}} \leq t^*$.
- Hence: $\bar{x} - t^* \frac{s}{\sqrt{n}} \leq \mu \leq \bar{x} + t^* \frac{s}{\sqrt{n}}$.
- In short: $\mu = \bar{x} \pm t^* \frac{s}{\sqrt{n}}$.

Statistical tests

Tests of significance

- Hypothesis: statement about popul. parameter.
- Result of test: probability P measuring how well data and hypothesis agree.
- **Null hypothesis** H_0: statement being tested ("no effect", "no difference").
- **Alternative hypothesis** H_a: what we aim at proving.

Significance of a test

- **P-value:** assuming H_0 is true, the probability that statistic take a value *as extreme as, or more extreme than* actually observed.

- Set statistical significance level α. If $P \leq \alpha$, then *data are statistically significant at level* α.

What to report

- Report conclusion ("alternative hypothesis true", or "data do not provide sufficient evidence to reject null hypothesis") at significance level α.

- Report also P-value (or: $P < 0.001$).

- Typically $\alpha = 0.05$. Stronger claims with lower α.

Dangers of using many tests

- Experiment repeated many time: by yourself or in different labs. (Cf. Bonferroni procedure.)

- Run experiment on different cases than pilot study (first observations motivating your experiment).

One-sided vs. two-sided

- Typically **null hypothesis**: $H_0 : \mu = \mu_0$.
- One-sided alternative hypothesis: $H_a : \mu > \mu_0$.
- One-sided alternative hypothesis: $H_a : \mu < \mu_0$.
- Two-sided alternative hypothesis: $H_a : \mu \neq \mu_0$.
- P-value doubled for two-sided.
- Two-sided, unless $a\ priori$ arguments for one-sided.

z-test

- Given data x_1, x_2,... x_n, and
- given **null hypothesis**: $H_0 : \mu = \mu_0$:
- 1. calculate $z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$,
- 2. use Table A or D to determine P-value, that is, the probability of sample mean being at least as extreme as $\bar{x}$.
- Proceed as on earlier chart.

t-test

- Given data x_1, x_2,... x_n, and
- given **null hypothesis**: $H_0 : \mu = \mu_0$:
- 1. calculate $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$,
- 2. use Table D or software to determine P-value, that is, the probability of sample mean being at least as extreme as $\bar{x}$.
- Proceed as on earlier chart.

Matched pairs t procedure

Read in M&M: good example, clearly presented.
If you get it, you have understood everything needed.
Relevant in experimental (psycho-)linguistic research.

Let me know if you have questions.

Two samples

One population vs. two populations

- So far: one population. Infer population mean μ from sample mean $\bar{x}$: which interval contains μ? can μ be equal to μ_0 of null hypothesis?

- Very often: two populations. Questions: are their means different? How much are their means different?

Two populations, two samples

- **Population 1**: mean μ_1, standard deviation σ_1.
- Draw a SRS: size n_1, mean $\bar{x}_1$, st. dev. s_1.
- **Population 2**: mean μ_2, standard deviation σ_2.
- Draw a SRS: size n_2, mean $\bar{x}_2$, st. dev. s_2.

Interested in $\mu_1 - \mu_2$. Infer it from $\bar{x}_1 - \bar{x}_2$.

A note

- We focus on the difference of their **means**.
- One can also ask about difference of other statistics. F-test (section 7.3): are standard deviations σ_1 and σ_2 different?
- Don't learn it, but remember where to find it.

Two-sample z-statistics

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

- Estimate confidence interval for $\mu_1 - \mu_2$.

- Null hypothesis: $\mu_1 = \mu_2$.

Two-sample t-statistics

$$t = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Estimate confidence interval for $\mu_1 - \mu_2$.
- Null hypothesis: $\mu_1 = \mu_2$.

How to use it?

- Nasty behavior... so
- approximate it using t-distribution $t(k)$, such that
- k is the smaller one of $n_1 - 1$ and $n_2 - 1$.
- Or, rather: use software.
- Best if $n_1 = n_2$ and large. But not always possible.

Next week:

- Proportions: applying all of this to a special case, called binomial distributions.