Statistics for EMCL week 2

Tamás Biró

Humanities Computing

University of Groningen

t.s.biro@rug.nl

This week:

Basics of inferential statistics:

- Sampling distribution (M&M 3.3)
- Central limit theorem (M&M 5.2)

  If interested in maths: chapters 4-5; not required.

- Normal distribution (M&M 1.3)
- Confidence intervals (M&M 6.1)
- Tests of significance; z-test (M&M 6.2-3).

Inferential statistics

- Examine **sample** drawn from the entire **population** in a carefully **designed** way (M&M 3.1-2; ethical issues: 3.4).

- **Parameter**: a number that describes population. A fixed, but unknown value. That is what interests us.

- **Statistic**: a number that describes the sample. Its value can be calculated from the results of the experiment, and used to estimate parameter. Changes from sample to sample.

$$
\begin{tikzcd}
population \arrow[r] \arrow[d] & sample \arrow[d] \\
parameter & statistic \arrow[l]
\end{tikzcd}
$$

Procedure of inference

- Sample $\longrightarrow$ statistic $\longrightarrow$ parameter estimated.

- Trustworthiness of procedure: what would happen if we repeat this procedure many times?

- Probability theory: field of mathematics describing the behavior of random processes, such as sampling. (Intro: M&M ch. 4.)

Sampling distribution

- **Sampling distribution of a statistic**: distribution of the statistic, if taken from all possible samples of population.

Example: measure height of people in a sample, then take the median $\times$ repeat it for many samples $\longrightarrow$ sample distribution of the medians.

Sampling distribution

- **Bias**: center of sampling distribution
  Unbiased statistic: mean of its distribution = true value of the parameter.
- **Variability of statistic**: spread of sampling distribution.

GOOD NEWS!

Sampling distribution

- To reduce bias: use random sampling.

  (Problem belonging to methodology and experiment design, not to mathematical statistics.)

- To reduce variability: use larger sample.

- Population size (if much larger than sample) does not matter.

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 $\frac{\sigma}{\sqrt{n}}$.

Normal (Gaussian) distribution

$$N(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

- $e = 2.7182\ldots$. Mean: μ. Standard deviation: σ.
- Area under curve is 1.
- 68–95–99.7 rule: area within $1/2/3$ σ from μ. See nice figures in M&M or JN's slides.

Standard Normal distribution

- $\mu = 0$ and $\sigma = 1$.
- Standardizing observations: $z = \frac{x-\mu}{\sigma}$
- **Cumulative proportions** of standard Normal distribution: area under the curve left to z: see Standard Normal Table.

Normal calculations, inverse Normal calculations

- Calculate area $right$ to $z = 1.47$.
- Find area from $z = -1.82$ to $z = 0.93$.
- What is z if left to it you find area 0.300?
- Similar questions with any other Normal distribution: normalize it ($x \to z$) first.

Normal calculations, inverse Normal calculations

And now, you:

- For what z is 95% of area between $-z$ and z?
- For what z is 5% of area right of z?

Normal quantile plots

Do data follow Normal distribution?

- Arrange observed data values from smallest to largest. Record what percentile a value occupies.
- Normal score: z value of percentile in st. Norm. distr. That is the value that the corresponding percentile should have had, if the distribution was really Normal.
- Plot data against corresponding Normal score.

If data follow Normal distribution, then plotted points lie close to a straight line.

Now back to inferences!

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.

Searching for population mean μ

- This week: we suppose standard deviation σ of population is known.
- Most often: unrealistic assumption.
- Next week: what if unknown? $\longrightarrow$ t-test.

Confidence interval

- Repeating data sampling many times: how often $\bar{x}$ between $\mu - m$ and $\mu + m$?
- Central Limit Theorem: first standardize, then use Standard Normal Table.
- Probability of $\bar{x}$ being between $\mu - m$ and $\mu + m$ is the same as unknown μ being between $\bar{x} - m$ and $\bar{x} + m$ for specific $\bar{x}$.

Confidence interval

- Set **confidence level** C: probability of interval containing true value of parameter (e.g., μ) is C.
- Use Standard Normal Table to find z^*:

z^*	1.645	1.960	2.576
C	90%	95%	99%

- **Margin of error**:

$$m = z^* \frac{\sigma}{\sqrt{n}}$$

where n = sample size; σ = known standard deviation in population.

- **Confidence interval**: $\bar{x} \pm m$.

We are confident at level C that the population mean (the unknown parameter) is between $\bar{x} - m$ and $\bar{x} + m$.

Remarks

- How conf. intervals behave: change C, n (σ).
- Choose sample size: $n = \left(\frac{z^*\sigma}{m}\right)^2$.
- Not Normal distribution in population (e.g., skewed distribution): only for large n (e.g., $n > 15$; use Normal quantile plots).
- Fancy mathematics cannot compensate for flaws in data collection (not randomize sample; outliers...).