Statistical inference (MVE155/MSG200)

Random sampling

Student representatives

- Rasmus Andersson (MPBME)
- Martin Ekerstedt (MPBME)
- Linnéa Fransson (MPSOF)
- Charlotte Fiona Preunkert (MPENM)
- Gabriel Shafiq Ahlgren (MPENM)

Topics

- Random sample: definition
- Point estimation
- Interval estimation
- Random sampling versus simple random sampling
- Stratified sampling

Population and random sample

- We have a population of interest and are interested in some property of it .
- Using probability theory, we can draw conclusions of a population based on only a sample, a subset of the population.
- A random sample of size n from the distribution of a random variable X is a collection of independent random variables that have the same distribution as X.
- Statistical inference is an estimate, prediction, or some other generalization of a large population that we make based on a random sample from the population.

Population and random sample

Let $X_1, ..., X_n$ be a vector of iid random variables, i.e. a random sample, and $x_1, ..., x_n$ a realization of it, i.e. sample.

Any function $g(x_1, ..., x_n)$ of the sample is called a statistic. For example, the sample mean and variance

$$\bar{x} = \frac{1}{n}(x_1 + ... + x_n) \text{ and } s^2 = \frac{1}{n-1}((x_1 - \bar{x})^2 + ... + (x_n - \bar{x})^2).$$

are statistics.

Point estimation

Let us have a population distribution of interest with parameter θ and we want to estimate θ based on the data (sample) $x_1, ..., x_n$.

We choose a relevant statistic $g(x_1, ..., x_n)$ as a point estimate for θ, i.e. $\hat{\theta} = g(x_1, ..., x_n)$. The corresponding random variable

$$\hat{\Theta} = g(X_1, ..., X_n)$$

is called a point estimator and the distribution of it is called the sampling distribution of the point estimator.

Point estimation

The quality of the point estimator can be measured by

- its expected value $\mathbb{E}(\hat{\Theta})$
- variance $\mathsf{Var}(\hat{\Theta})$
- and/or their combination, the mean square error

$$\mathbb{E}((\hat{\Theta} - \theta)^2) = (\mathbb{E}(\hat{\Theta}) - \theta)^2 + \mathsf{Var}(\hat{\Theta})$$

where the bias $\mathbb{E}(\hat{\Theta}) - \theta$ measures the lack of accuracy (systematic error) and the variance $\mathsf{Var}(\hat{\Theta})$ the lack of precision (random error).

Point estimation: unbiasedness and consistency

If the bias is zero, or $\mathbb{E}(\hat{\Theta}) = \theta$, the estimator is unbiased.

The estimator is consistent if the mean square error

$$\mathbb{E}((\hat{\Theta} - \theta)^2) = (\mathbb{E}(\hat{\Theta}) - \theta)^2 + \mathrm{Var}(\hat{\Theta})$$

vanishes as $n \to \infty$, i.e. that

- the estimator is asymptotically unbiased and
- the variance of the estimator vanishes $n \to \infty$.

This means that a consistent estimator $\hat{\Theta}$ approaches the true parameter value.

The sample mean $\bar{X}$ and the sample variance S^2 are unbiased and consistent estimators of the population mean μ and the population variance σ^2, respectively:

$$\mathbb{E}[\bar{X}] = \mu, \quad \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$$

and

$$\mathbb{E}[S^2] = \sigma^2, \quad \mathrm{Var}(S^2) = \frac{\sigma^4}{n}\left(\mathbb{E}\left[\left(\frac{X-\mu}{\sigma}\right)^4\right] - \frac{n-3}{n-1}\right).$$

Note that the sample standard deviation S is not an unbiased estimator for the population standard deviation σ.

Standard error

The standard deviation of the estimator $\hat{\Theta}$,

$$\sigma_{\hat{\Theta}} = \sqrt{\mathrm{Var}(\hat{\Theta})}$$

is called the standard error of the point estimate. The estimated standard error $s_{\hat{\theta}}$ of the point estimate is a point estimate of $\sigma_{\hat{\Theta}}$ computed from the data.

The standard error of the sample mean is estimated by $s_{\bar{x}} = s/\sqrt{n}$.

Interval estimation: approximate confidence interval for μ

If the sample size n is large enough,

$$\bar{X} \approx N(\mu, \sigma)$$

and

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx N(0, 1)$$

independently of which distribution the observations come from.

Furthermore, since S is a consistent estimator for σ,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \approx N(0, 1).$$

Interval estimation: approximate confidence interval for μ

A $100(1-\alpha)\%$ confidence interval for μ can be approximated by using normal distribution:

$$I_\mu \approx \bar{x} \pm z(\alpha/2)s/\sqrt{n} = \bar{x} \pm z(\alpha/2)s_{\bar{x}},$$

where $\Phi(z(\alpha)) = 1-\alpha$, $\alpha \in (0,1)$.

It can be seen that

- the higher the confidence level, the wider the confidence interval
- the larger the sample variance, the wider the confidence interval
- the larger the sample size n, the narrower the interval.

Interval estimation: exact confidence interval for μ

If the sample size n is small

- $\bar{X}$ is (approximatively) normal only if the sample $x_1, ..., x_n$ comes from a normal distribution.
- σ cannot be replaced by S.

Given that the sample comes from a normal distribution, i.e. $X_i \sim N(\mu, \sigma)$, $i = 1, .., n$, an (exact) $100(1-\alpha)\%$ confidence interval for μ can be computed by using the t-distribution since

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{n-1}.$$

The confidence interval for μ becomes

$$I_\mu = \bar{x} \pm t_{n-1}(\alpha/2)s/\sqrt{n} = \bar{x} \pm t(\alpha/2)s_{\bar{x}},$$

where $t_{n-1}(\alpha)$ is defined similarly to $z(\alpha)$.

Interval estimation: exact confidence interval for σ^2

We saw earlier that if the observations come from $N(\mu, \sigma)$, then

$$\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$$

and

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}.$$

A $100(1-\alpha)\%$ confidence interval for σ^2 is then

$$I_{\sigma^2} = \left(\frac{(n-1)s^2}{x_{n-1}(\alpha/2)}, \frac{(n-1)s^2}{x_{n-1}(1-\alpha/2)} \right),$$

where $x_{n-1}(\alpha)$ is defined similarly to $z(\alpha)$ and can be found in the χ^2 table.

Dichotomous data

In dichotomous data, only two values 0 and 1 are possible, e.g. "heads" and "tails" in a coin toss if we convert the data as heads$=1$ and tails$=0$. In such a case, the Bernoulli distribution (for the outcome X) with parameter

$$p = \mathrm{P}(X = 1)$$

can be used as the population distribution.

Then, $\mu = p$ and the sample mean $\bar{x}$ is the same as the sample proportion $\hat{p}$. The sample proportion is an unbiased and consistent estimator for p. The standard error for the sample proportion can be estimated by

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1 - \hat{p})}{n - 1}}.$$

When the sample size is large, approximative confidence intervals can be estimated using the normal approximation and the estimated standard error above.

Simple random sampling

A finite population of size N can be thought as a set of N elements characterized by their numerical values $x \in \{a_1, ..., a_N\}$. Then, the population distribution is

$$P(X = x) = \frac{N_x}{N},$$

where N_x is the number of elements with $a_i = x$.

Random samples from a finite population can be taken

1. with replacement resulting in a random sample consisting of independent and identically distributed observations.
2. without replacement resulting in a simple random sample consisting of identically distributed but dependent observations.

When the sample size n is small compared to the population size N (less than 5% of the population), the two approaches are almost the same.

Simple random sampling

In the case of simple random sample $(X_1, ..., X_n)$ with dependent observations, the sample mean $\bar{X}$ is an unbiased and consistent estimator for the population mean with

$$\mathbb{E}(\bar{X}) = \mu \text{ and } \mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}(1 - \frac{n-1}{N-1})$$

and $1 - \frac{n-1}{N-1} = \frac{N-n}{N-1}$ is called the finite population correction.

Simple random sampling: sample variance

The sample variance S^2 is a biased estimator for the population variance σ^2 in this case since

$$\mathbb{E}(S^2) = \sigma^2 \frac{N}{N-1}.$$

Replacing σ^2 by $\frac{N-1}{N} S^2$ in the formula for $\mathrm{Var}(\bar{X})$, we obtain an unbiased estimator for $\mathrm{Var}(\bar{X})$, namely

$$S_{\bar{X}}^2 = \frac{S^2}{n}(1 - \frac{n}{N}).$$

Simple random sampling

If we have a rather large sample (more than 5% of the population) and use simple random sampling (without replacement), the corrected estimator for the variance should be used.

For example, for dichotomous data, the standard error becomes

$$s_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n-1}}\sqrt{1-\frac{n}{N}}$$

which will be used e.g. when constructing confidence intervals.

Stratified random sampling

Additional information on the population structure can be used to reduce the sampling error
$\rightarrow$ stratified sampling.

The total population is divided into k strata. For example, the population of Swedish school children is divided into four strata: southern Sweden, western Sweden, eastern Sweden, and northern Sweden.

The total population size is N and it consists of k strata sizes $N_1, ..., N_k$ such that $N = N_1 + ... + N_k$. The strata fractions $w_i = N_i/N$, $i = 1, ..., k$ are assumed to be known.

Stratified random sampling: mean and variance

Given the (unknown) strata means and standard deviations μ_i and σ_i, respectively, the population mean and variance become

$$\mu = \sum_{i=1}^{k} w_i \mu_i \text{ and } \sigma^2 = \overline{\sigma^2} + \sum_{i=1}^{k} w_i (\mu_j - \mu)^2,$$

where $\overline{\sigma^2} = \sum_{i=1}^{k} w_i \sigma_i^2$ and $w_i + ... + w_k = 1$.

Stratified random sampling: estimation of the population mean

Take k independent samples, one from each strata, with sample sizes $n_1, \dots, n_k$ and compute the sample means $\bar{x}_1, \dots, \bar{x}_k$. Then, the stratified sample mean is

$$\bar{x}_s = \sum_{i=1}^{k} w_i \bar{x}_i.$$

which is an unbiased estimate for μ.

Stratified random sampling: estimation of the variance of the sample mean

The variance of $\bar{X}_s$ is

$$\text{Var}(\bar{X}_s) = \sum_{i=1}^{k} \text{Var}(w_i \bar{X}_i) = \sum_{i=1}^{k} w_i^2 \text{Var}(\bar{X}_i) = \sum_{i=1}^{k} \frac{w_i^2 \sigma_i^2}{n_i}$$

and can be estimated by

$$\sum_{i=1}^{k} \frac{w_i^2 s_i^2}{n_i},$$

where s_i is the sample standard deviation for strata i.

Stratified random sampling: optimal allocation

We have n observations from the population of size N using stratified sampling, where n is much smaller than N (random sampling and simple random sampling almost the same).

What is the allocation $n_1, ..., n_k$ of the n observations that minimises the standard error $s_{\bar{x}}$ of $\bar{x}$?

The allocation, where

$$n_i = n \frac{w_i \sigma_i}{\bar{\sigma}}$$

and $\bar{\sigma} = w_i \sigma_1 + ... + w_k \sigma_k$ gives the smallest error, namely

$$\text{Var}(\bar{X}_{so}) = \frac{(\bar{\sigma})^2}{n}$$

where $\bar{X}_{so}$ refers to the mean using the optimal allocation.

Stratified random sampling: proportional allocation

Since σ_i's are often unknown, the observations are often allocated proportionally to the strata sizes so that $n_i = nw_i$, $i = 1, ..., k$.

This gives the usual sample mean $\bar{x}$ but a slightly larger variance

$$\mathrm{Var}(\bar{X}_{sp}) = \frac{\overline{\sigma^2}}{n},$$

where $\overline{\sigma^2} = w_1\sigma_1^2 + ... + w_k\sigma_k^2$.

Comparison

Sample means and sample variances:

	Sample mean	Variance of sample mean
Random sample	$\bar{x}$	$\frac{\sigma^2}{n}$
Stratified optimal	$\bar{x}_{so} = \sum_{i=1}^{k} w_i \bar{x}_i$	$\frac{(\bar{\sigma})^2}{n}$
Stratified proportional	$\bar{x}_{sp} = \bar{x}$	$\frac{\overline{\sigma^2}}{n}$

where $\bar{\sigma} = w_1\sigma_1 + ... + w_k\sigma_k$, $\overline{\sigma^2} = w_1\sigma_1^2 + ... + w_k\sigma_k^2$, and

$$\frac{(\bar{\sigma})^2}{n} \leq \frac{\overline{\sigma^2}}{n} \leq \frac{\sigma^2}{n}.$$