Lecture 1 – Program

1. Introduction

2. Probability concepts

3. Mean squared errors

4. Confidence intervals and hypothesis tests

5. Robustness and rank tests

(We skip Monte Carlo tests, pp. 8-10 in B&S)

Basic idea

The basic idea for the development and evaluation of most methods in statistics is to consider the data as generated by a probability model, and judge the variability of the data actually observed in relation to data generated from the probability model.

Thus one has:

- Actual empirical data, the sample, which is often described using numerical measures such as the mean and the standard deviation

- A probability model describing the distribution of the data, from which one can infer the distribution of the numerical measures used to summarize the empirical observations.

Example (B& S, page 1)

Age of mineral samples (million years)						
249	254	243	268	253	269	287
241	273	306	303	280	260	256
278	344	304	283	310		

Here we can compute the (empirical) mean, median and standard deviation:

$\bar{x} = 276.9, \quad \text{med} = 273.0, \quad s = 27.1$

In general we consider observations $x_1, \ldots, x_n$ that are either:

- replications of the same measurement (as in the example)

or

- observations on a random sample from some population

Observations may be quantitative (numeric) as in the example above, or qualitative (categoric). We will focus on quantitative data in the first part of the course.

Random variables and distributions

Observations (measurements) can be more or less variable (precise).
To describe the variability, we consider the data as independent replications of a random variables X, having a distribution described by a *probability density*, $f(x)$, or a *cumulative distribution function*, $F(x)$.

It is not possible to predict one realization of X, but it is possible compute the probability that it falls in a certain interval:

$$P(a < X \leq b) = \int_a^b f(x)\,dx = F(b) - F(a)$$

Distributions are described by (theoretical) summaries such as

- Mean or expectation: $\mu = E(X) = \int_{-\infty}^{\infty} x f(x) dx$
- Variance: $\sigma^2 = \mathrm{Var}(X) = \int_{-\infty}^{\infty} (x-\mu)^2 f(x) dx$
- Standard deviation: $\sigma = \mathrm{stan}(X) = \sqrt{\mathrm{Var}(X)}$

(The formulas above apply for a continuously distributed random variable. Similar formulas with sums apply for discrete random variables, e.g., counts.)

Law of large numbers

It is a common experience that empirical means (i.e. averages) become more precise as the number of observations increases.

This empirical phenomenon has a mathematical counterpart:

If $x_1, \ldots, x_n$ are independent replications of a random variable X with expectation μ and variance σ^2, then

- $\bar{x}_n = \frac{1}{n} \sum_{i=1}^{n} x_i \to \mu$
- $s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \to \sigma^2$

as n increases.

The summation sign $\sum_{i=1}^{n}$ means that we should put $i = 1, 2, \ldots, n$ in the expression following the summation sign and add together the n terms thus obtained.

Central limit theorem

An implication of the law of large numbers is that if $x_1, \ldots, x_n$ are independent replications of a random variable X with expectation μ, then $\bar{x}_n - \mu \to 0$.

But if we "blow up" $\bar{x}_n - \mu$ at the right rate, the magnified difference will converge, not toward a number, but *in the sense that the distribution of the magnified* $\bar{x}_n - \mu$ *looks more and more normal*. This is the central limit theorem.

What is the rate?

Remember that:

$$\mathrm{Var}(\bar{x}_n - \mu) = \mathrm{Var}(\bar{x}_n) = \frac{\sigma^2}{n}$$

This gives:

$$n\,\mathrm{Var}(\bar{x}_n - \mu) = \mathrm{Var}\left(\sqrt{n}(\bar{x}_n - \mu)\right) = \sigma^2$$

This indicates that the right rate is $\sqrt{n}$

Central limit theorem, contd.

Notation:

- $Z \sim N(0,1)$ means that Z is a standard normal random variable.
  (Note that $E(Z) = 0$ and $\text{Var}(Z) = 1$.)

- $\Phi(z) = P(Z \leq z)$ is the cumulative distribution function of a standard normal random variable.

The mathematical formulation of the central limit theorem is:

If $x_1, \ldots, x_n$ are independent replications of a random variable X with expectation μ and variance σ^2, then

$$P\left(\frac{\sqrt{n}(\bar{x}_n - \mu)}{\sigma} \leq z\right) \to \Phi(z)$$

as n increases.

Mean squared error

The purpose of an investigation is often to *estimate* an unknown quantity. This may be a *parameter* describing the probability model, or a function of the model parameters.

To be specific, let us consider the situation where the empirical mean $\bar{x}_n$ is used to estimate a quantity q (e.g. the median, m, of the population distribution, $F(m) = 1/2$).

The quality of the estimation is judged by the error

$$e_n = \bar{x}_n - q$$

Note that the estimation error is a random quantity.

The *mean squared error* is defined as $E(e_n^2)$, and it is the squared estimation error "in the long run".

Mean squared error, contd.

One may show that

$$E(e_n^2) = \frac{\sigma^2}{n} + (\mu - q)^2,$$

where $E(\bar{x}_n) = \mu$.

This means that the mean squared error can be decomposed in two parts: one due to randomness, which vanishes when the number of observations increase, and a (squared) *bias* term which is due to systematic errors and does not vanish.

If we have $q = \mu$, then $\bar{x}_n$ is *unbiased*, and the mean square error equals $\mathrm{Var}(\bar{x}_n) = \sigma^2/n$.

When a normal approximation is appropriate, the quality of the estimation may be assessed by the $68-95-99.7$ rule, e.g. 68% of all replicated estimations will fall within $\pm\sigma/\sqrt{n}$ of the true unknown value μ.

Confidence intervals

The typical form of a confidence interval is

$$\text{estimate} \pm c \cdot se(\text{estimate}).$$

where $se(\text{estimate}) = \sqrt{\text{Var(estimate)}}$ is the *standard error* of the estimate (and usually has to be estimated, cf. below).

In general a confidence interval (c.i.) for an unknown quantity q has the form (a, b), where a and b are computed from the data.

The *confidence coefficient* $1 - \alpha$ of a c.i. is the probability that the interval contains the unknown quantity:

$$P(a < q < b) = 1 - \alpha$$

Example, c.i. for the mean

Suppose that $x_1, \ldots, _n$ is a random sample from $N(\mu, \sigma^2)$ (i.e. the normal distribution with mean μ and variance σ^2)

(i) σ known

$\bar{x}_n \sim N(\mu, \sigma^2/n)$, and a confidence interval takes the form (cf. above):

$$\bar{x}_n \pm c \cdot \frac{\sigma}{\sqrt{n}}$$

c is defined (implicitly) by

$$P(\bar{x}_n - c \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x}_n + c \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

One may find the values of c from a table of the standard normal distribution. In particular one finds the following values of c:

$1-\alpha$	90%	95%	99%
c	1.645	1.960	2.576

(ii) σ **unknown**

When σ^2 is unknown (as is usually the case), we may estimate σ by the empirical standard deviation s_n.

A confidence interval then takes the form:

$$\bar{x}_n \pm c \cdot \frac{s_n}{\sqrt{n}}$$

We now have to use the *t-distribution with n-1 degrees of freedom* to determine c.

As n grows larger, the quantiles of the t-distribution(s) approach those of the standard normal, cf. Table 3 page 7 in B&S.

Example, mineral samples

A 95% c.i. has limits

$$276.9 - 2.10 \times 27.1/\sqrt{19} = 263.8$$

and

$$276.9 + 2.10 \times 27.1/\sqrt{19} = 290.0$$

Hypothesis testing

Setup:

- Assume that $H_0 : \mu \leq \mu_0$ denotes a set of values of interest
- We observe $\bar{x}_n$ and s_n in a random sample from a normal population
- Can we reject the null hypothesis H_0?

This is usually done through the *p-value*.

The p-value is the probability that the *test statistic* has a value equal to or more "extreme" than the one observed *when H_0 is true.* In other words we compute the evidence *against H_0*.

Example, test of the mean

Again we have two situations:

(i) σ **known**
We reject H_0 for large values of the test statistic

$$z = \frac{\bar{x}_n - \mu_0}{\sigma/\sqrt{n}}$$

Under H_0 the test statistic is standard normally distributed, and that can be used to compute the p-value: $p = P(Z > z)$.

(ii) σ **unknown**
We reject H_0 for large values of the test statistic

$$t = \frac{\bar{x}_n - \mu_0}{s_n/\sqrt{n}}$$

Under H_0 the test statistic is t-distributed with n-1 degrees of freedom, and that can be used to compute the p-value: $p = P(T_{n-1} > t)$.

Mineral samples

We have $\bar{x_n} = 276.9,\ s_n = 27.1,\ \mu_0 = 265$.

We have

$$t = \frac{276.9 - 265}{27.1/\sqrt{19}} = 1.91$$

This corresponds to a p-value of 3%.

Therefore, it is not plausible that the area where the mineral samples were collected, is less than $\mu_0 = 265$ million old.

Robustness

Robustness refers to investigations of how sensitive a method is to the underlying assumptions for its validity, and also to the search for methods that cover a broad spectrum of population models.

It is also important to investigate to *which deviations* a method is sensitive.

In B&S a small investigation is carried out for a situation corresponding to the mineral samples.

The population model is

$$x_i = \mu + \sigma\varepsilon_i \quad i = 1, \cdots, 19$$

where $\varepsilon_1, \ldots, \varepsilon_{19}$ are independent and either

- standard normal, or
- t-distributed with 5 degrees of freedom, or
- having density $f(x) = \exp(-(x+1))$ for $x > 1$.

The first and second distributions are symmetric, the third is skewed to the right.

Confidence intervals of the form

$$\bar{X}_n \pm c\frac{s_n}{\sqrt{n}}$$

were computed for the mineral data for each of the three population models (the appropriate c-values were found by simulation for the latter two population models)

From Table 5, page 11 in B&S, we see that the c.i's based on the t_5-distribution are very close to the ones based on the normal distribution, while the one based on the last distribution is shifted a bit (but not much) to the left.

In conclusion: Confidence intervals (and tests) for the mean μ are quite robust to distributional assumptions. This is due to the central limit theorem.

B&S also study confidence intervals and tests for the variance σ^2.

These depend a lot on the assumed population distribution; cf. Tables 5 and 6, page 11 in B&S.

In conclusion: Confidence intervals (and tests) for the variance σ^2 are *not* robust to distributional assumptions.

Rank-based methods

These methods are valid for a broad class of population distributions.

Wilcoxon's signed rank test

Consider the null hypothesis $H_0 : m \leq 265$ for the mineral sample, where m is the median given by $F(m) = 1/2$.

Wilcoxon's signed rank test is a nonparametric test for this hypothesis, valid for all symmertic population distributions

The test statistic is computed as follows:

- Subtract 265 from all observations, x_i
- rank or sort $|x_i - 265|$
- compute the sum, T^+_{obs}, of the ranks corresponding to observations with $x_i - 265 > 0$.
- For small n: The distribution of T^+ is tabulated. Reject H_0 if T^+_{obs} large.
- For large n: Use that the distribution of $Z = (T^+ - (n(n+1)/4)/\sqrt{n(n+1)(2n+1)/24}$ is approximately standard normal when $F(m) = 1/2$.

x_i	249	254	243	268	253	269	287
	241	273	306	303	280	260	256
	278	344	304	283	310		
$x_i - 265$	-14	-11	-22	3	-12	4	22
	-24	8	41	38	15	-5	-9
	13	79	39	18	45		
Rank	9	6	12.5	1	7	2	12.5
	14	4	17	15	10	3	5
	8	19	16	11	18		
Sign	-	-	-	+	-	+	+
	-	+	+	+	+	-	-
	+	+	+	+	+		

In this case $T^+_{obs} = 133.5$. From a table $P(T^+ \leq 133) = 0.938$.

Also, $z = 1.55$, and $P(Z > 1.55) = 0.061$, so the normal approximation is good.

Note, that we weigh the $x_i - 265 > 0$ by the ranks. If instead, we only count the number of $x_i - 265 > 0$, we get the so-called *Sign test*. The critical values and p-values are given by the $Bin(n, 1/2)$ distribution. In this case, 12 of the $x_i - 265$'s are positive. Since $P(X \leq 11) = 0.8204$, this test has a p-value of 18%.