STK4900/9900 - Lecture 1

Program

1. Introduction
2. Descriptive methods
3. Data and probability models
4. Normal distributions
5. Estimation and confidence intervals
6. Hypothesis testing and P-values
7. Robustness
8. Bootstrapping

- Sections 2.1 – 2.3
- Sections 3.1.1 – 3.1.3, 3.1.7 and 3.6
- Supplementary material

  (cf. your introductory statistics textbook)

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 to judge the variability of the data actually observed in relation to data generated from that 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

Descriptive methods

Measure the age of 19 mineral samples from the Black Forest in Germany using potassium-argon dating

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		

To summarize the data we may compute (e.g.) the (empirical) *mean*, *median* and *standard deviation*:

$$\bar{x} = 276.9$$

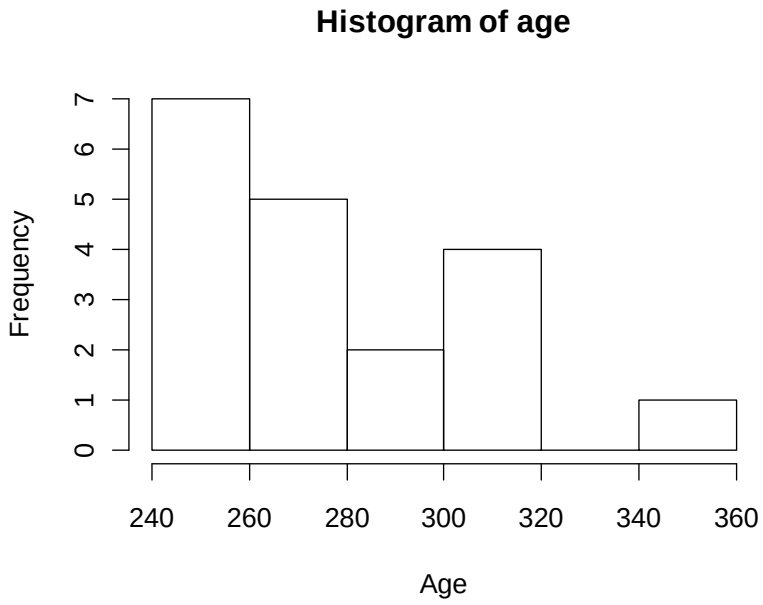
$$\text{med} = 273.0$$

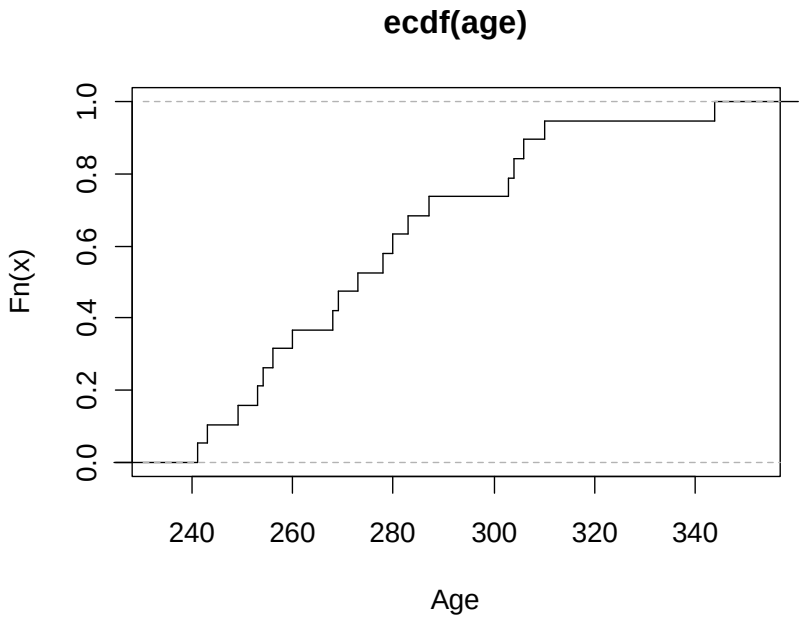
$$s = 27.1$$

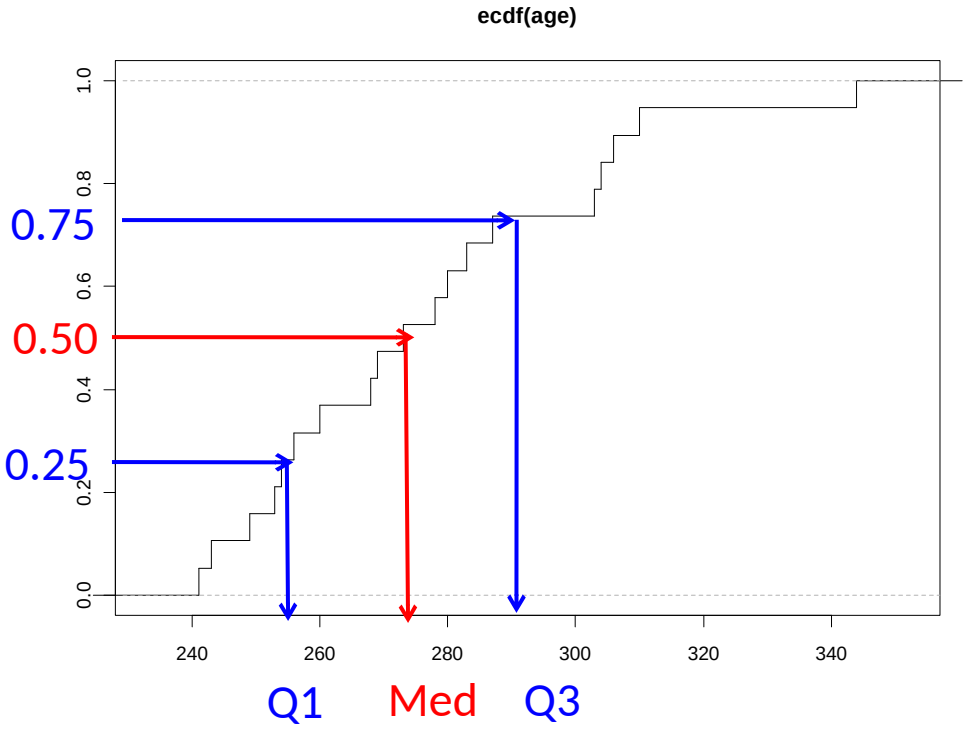
(Formulas for $\bar{x}$ and s are given below)

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		

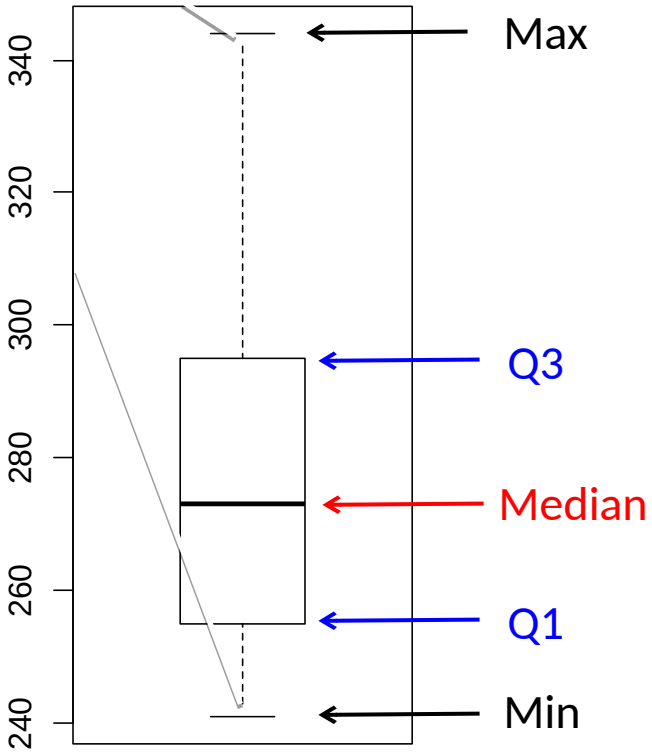
The distribution of the data may be illustrated (e.g) by a *histogram* and by the *empirical cumulative distribution* (ecdf)







A *boxplot* gives another useful graphical display for a data set:



Other summary measures are the *first quartile* (Q1) and the *third quartile* (Q3)

When we use statistical software (like R) to compute the quartiles, the software may adopt some interpolations which make the values of the quartiles differ somewhat from those read directly from the ecdf

Data

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 *numerical* (as in the example) or *categorical* (e.g. gender)

We focus on numerical data in the first part of the course

Empirical mean and standard deviation for numerical data:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

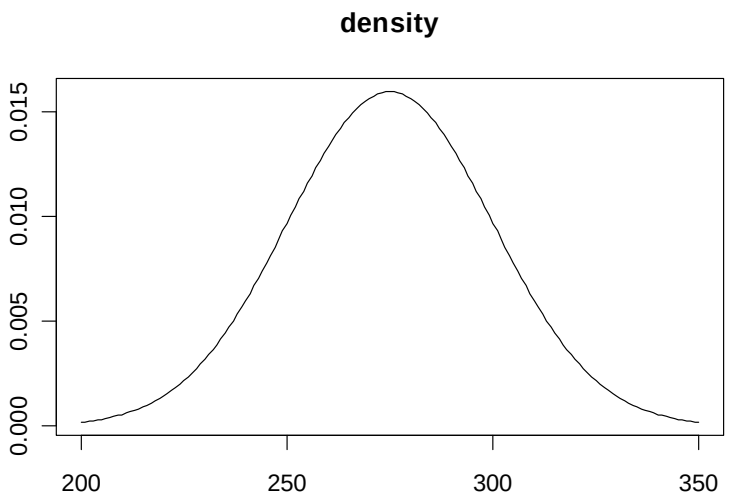
Random variables and distributions

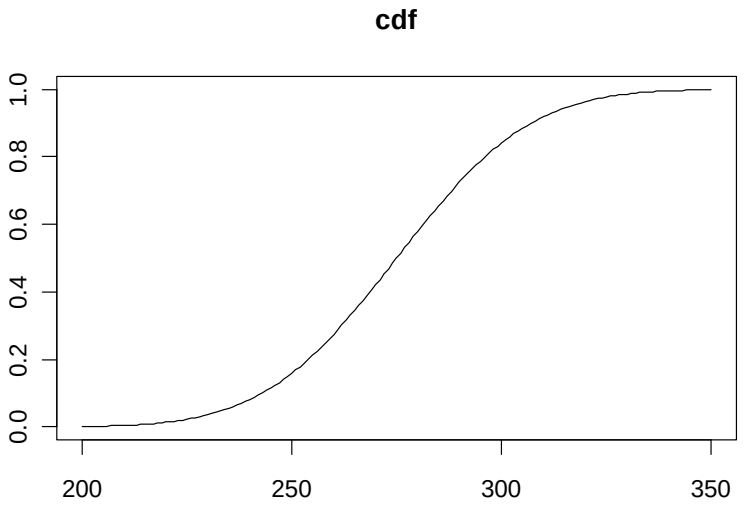
Observations (measurements) can be more or less variable (precise)

To describe the variability, we consider the data as independent replications of a random variable X, having a distribution described by a *probability density* $f(x)$ or a *cumulative distribution function* (cdf) $F(x)$

$f(x)$ and $F(x)$ are the theoretical counterparts to the histogram and the ecdf, respectively

Example: Density and cdf for normally distributed random variable with (theoretical) mean 275 and standard deviation 25

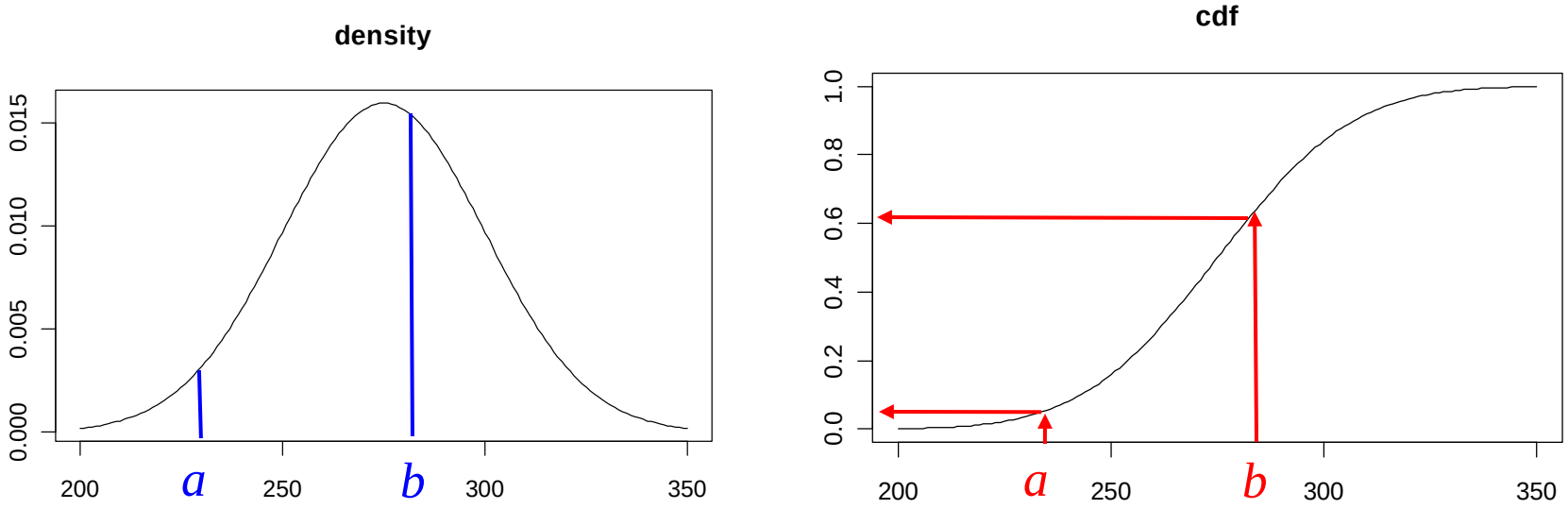




It is not possible to predict exactly one realization of X, but it is possible compute the probability that it falls in a certain interval:

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

Illustration:



In practice statistical tables or statistical software (like R) are used to find the probabilities

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:* $\sigma = \mathrm{sd}(X) = \sqrt{\mathrm{Var}(X)}$

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

Properties of expectation and variance:

$$E(a+bX) = a + bE(X)$$
$$E(X+Y) = E(X) + E(Y)$$

$$\mathrm{Var}(a+bX) = b^2 \mathrm{Var}(X)$$
$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) \quad \text{when } X \text{ and } Y \text{ are independent}$$

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 in the law of large numbers:

Suppose that $x_1, \ldots, x_n$ are independent replications of a random variable X with mean μ and standard deviation σ, then

- $\overline{x} \rightarrow \mu$

as n increases

One also has that $s \rightarrow \sigma$ as n increases

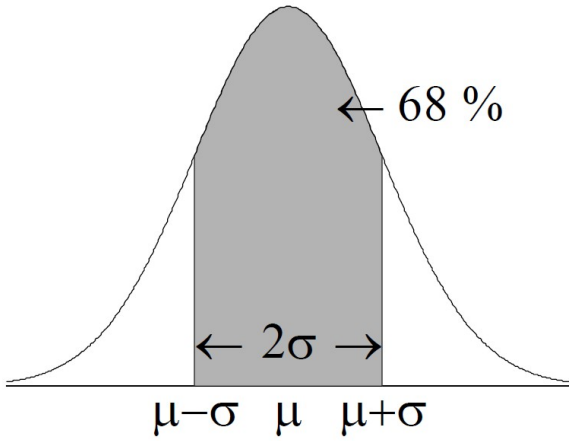
Normal distributions

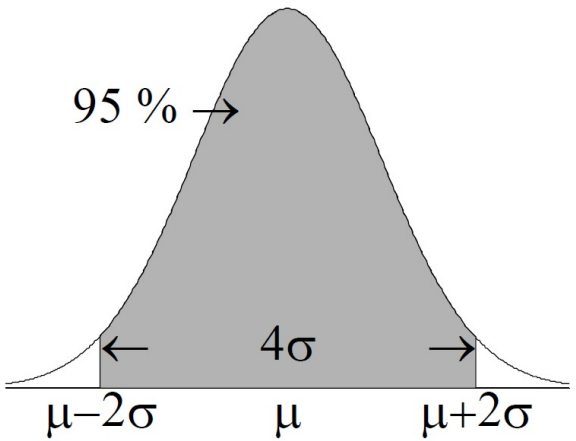
The normal distributions will play a key role in the first part of the course

A random variable X is normally distributed with mean μ and standard deviation σ [short: $X \sim N(\mu, \sigma^2)$] if its density takes the form:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \cdot \left(\frac{x - \mu}{\sigma} \right)^2 \right\}$$

For all normal distributions we have:





If a random variable Z is normally distributed with mean $\mu = 0$ and standard deviation $\sigma = 1$, we say that Z is standard normally distributed, i.e. $Z \sim N(0,1)$

Two important results:

1) If $X \sim N(\mu, \sigma^2)$, then $Z = \frac{X - \mu}{\sigma} \sim N(0,1)$

2) Suppose that $x_1, \ldots, x_n$ are independent replications of a random variable $X \sim N(\mu, \sigma^2)$,
[we say that $x_1, \ldots, x_n$ is a random sample from the normal distribution with mean μ and standard deviation σ],
then $\bar{x} \sim N(\mu, \sigma^2 / n)$

If the sample size is reasonably large, result 2 holds approximately also when $x_1, \ldots, x_n$ is a random sample from another distribution than the normal (central limit theorem)

Estimation

The purpose of an investigation is often to use the data to *estimate* an unknown quantity θ

θ may be a parameter describing the probability model (e.g. the mean μ or the standard deviation σ), or a function of the model parameters (e.g. the coefficient of variation σ/μ)

To be specific, consider the situation where the empirical mean $\overline{x}$ is used to estimate the mean μ of a distribution

It is then common to write $\hat{\mu} = \overline{x}$

In the example with mineral samples, we estimate the age to be $\hat{\mu} = \overline{x} = 276.9$ million years

In general we have an *estimator* $\hat{\theta}$ for the unknown θ

Note that $\hat{\theta}$ is a random variable (since it depends on the data), and hence we may consider the expected value and the variance of $\hat{\theta}$

We will only consider estimators that are *unbiased*, i.e. $E(\hat{\theta}) = \theta$, (or that are approximately unbiased) i.e. estimators that give (approximately) the correct value "in the long run"

Then the uncertainty of an estimator may be measured by its *standard error* $se(\hat{\theta}) = \sqrt{\mathrm{Var}(\hat{\theta})}$

Consider the estimator $\hat{\mu} = \bar{x}$

Its standard error is given by $se(\bar{x}) = \sigma / \sqrt{n}$

In practice the standard error has to be estimated by replacing σ by the empirical standard deviation s

In the example with mineral samples, the (estimated) standard error becomes $s / \sqrt{n} = 27.1 / \sqrt{19} = 6.2$ million years

Confidence intervals

The typical form of a confidence interval is

$$\left(\hat{\theta} - c \cdot se(\hat{\theta}) \ , \ \hat{\theta} + c \cdot se(\hat{\theta})\right) \qquad (*)$$

where $se(\hat{\theta})$ is the (estimated) standard error of the estimate

In general a confidence interval for an unknown quantity θ has the form (L, U), where L and U are computed from the data.

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

$$P(L < \theta < U) = 1 - \alpha$$

The confidence interval (*) may more briefly be given as

$$\hat{\theta} \pm c \cdot se(\hat{\theta})$$

Confidence interval for the mean μ

Suppose that $x_1, \ldots, x_n$ is a random sample from $N(\mu, \sigma^2)$

σ known

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

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

c is defined (implicitly) by

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

One may find c from a table of the standard normal distribution

In particular for a 95% CI we have $c = 1.96$

σ unknown

When σ is unknown (as is usually the case), we may estimate σ by the empirical standard deviation s

A confidence interval then takes the form:

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

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

For the example with mineral samples a 95% CI 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

General set-up:

- We want to test the null hypothesis $H_0 : \theta \leq \theta_0$

  versus the (one-sided) alternative hypothesis $H_A : \theta > \theta_0$

- From the observed data we compute a *test statistic*

- Based on the observed value of the test statistic, can we reject H_0

  (and hence conclude that H_A is true)?

This is usually done through the *P-value*

The P-value is the probability, *when* H_0 *is true,* that the test statistic has a value equal to or more "extreme" than the one observed

In other words we compute the evidence against H_0 (i.e. in favor of H_A).

Test for the mean μ

Suppose that $x_1, \ldots, x_n$ is a random sample from $N(\mu, \sigma^2)$

We want to test the null hypothesis $H_0 : \mu \leq \mu_0$ versus the alternative $H_A : \mu > \mu_0$

Again there are two situations:

σ known

We reject H_0 for large values of the test statistic

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

Under H_0 the test statistic is standard normally distributed

That can be used to compute the (one-sided) P-value: $P = P(Z > z)$

where $Z \sim N(0,1)$

σ unknown

We reject H_0 for large values of the test statistic

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

Under H_0 the test statistic is t-distributed with n -1 degrees of freedom (df)

That can be used to compute the (one-sided) P-value: $P = P(T > t)$

where T is t-distributed with n -1 df.

Consider for illustration the example with mineral samples

We want to test the null hypothesis $H_0 : \mu \leq 265$ versus the alternative $H_A : \mu > 265$

We have $\bar{x} = 276.9,\ s = 27.1,\ \mu_0 = 265$

This gives

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

corresponding to a P-value of 3.6%

Therefore, we may reject the null hypothesis and conclude that the area where the mineral samples were collected is older than $\mu_0 = 265$ million years.

To be aware of: Lately, there have been extensive discussions in scientific journals and media about the misuse of P-values that has developed in some fields – originating from misconceptions of what P-values mean and how they should be used.

The ASA's Statement on p-Values: Context, Process, and Purpose

The American Statistician 2016 Volume 70(2) pp. 129-133

The statement's six principles, many of which address misconceptions and misuse of the p- value, are the following:

1. P-values can indicate how incompatible the data are with a specified statistical model.

2. P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.

3. Scientific conclusions and business or policy decisions should not be based only on whether a p-value passes a specific threshold.

4. Proper inference requires full reporting and transparency.

5. A p-value, or statistical significance, does not measure the size of an effect or the importance of a result.

6. By itself, a p-value does not provide a good measure of evidence regarding a model or hypothesis.

Comparing two groups

Measure bone mineral density (in g/cm^2) for rats given isoflavone and for rats in a control group:

1) Control ($n_1 = 15$)				
0.228	0.207	0.234	0.220	0.217
0.228	0.209	0.221	0.204	0.220
0.203	0.219	0.218	0.245	0.210
2) Isoflavone ($n_2 = 15$)				
0.250	0.237	0.217	0.206	0.247
0.228	0.245	0.232	0.267	0.261
0.221	0.219	0.232	0.209	0.255

Question: Does isoflavone have an effect on bone mineral density?

Means and standard deviations:

$$\bar{x}_1 = 0.2189 \qquad s_1 = 0.0116$$

$$\bar{x}_2 = 0.2351 \qquad s_2 = 0.0188$$

Suppose that the data for the two groups are random samples from $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$, respectively

We estimate $\mu_2 - \mu_1$ by the difference in the (empirical) means, i.e. by $\bar{x}_2 - \bar{x}_1$

Standard error (estimated): $$se(\bar{x}_2 - \bar{x}_1) = s_p \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

Here $$s_p = \sqrt{\frac{n_1 - 1}{n_1 + n_2 - 2} s_1^2 + \frac{n_2 - 1}{n_1 + n_2 - 2} s_2^2}$$

95% confidence interval for $\mu_2 - \mu_1$:

$$\bar{x}_2 - \bar{x}_1 \pm c \cdot se(\bar{x}_2 - \bar{x}_1)$$

where *c* is the upper 97.5% percentile in the t-distribution with $n_1 + n_2 - 2$ df

In the example the estimated effect of the treatment becomes:

$$\bar{x}_2 - \bar{x}_1 = 0.2351 - 0.2189 = 0.0162$$

Standard error:

$$se(\bar{x}_2 - \bar{x}_1) = 0.0156\sqrt{\frac{1}{15} + \frac{1}{15}} = 0.0057$$

95% confidence interval:

$$0.0162 \pm 2.05 \cdot 0.0057$$

i.e.

$$0.0162 \pm 0.0117$$

We then consider testing the null hypothesis $H_0 : \mu_1 = \mu_2$

versus the (two-sided) alternative $H_A : \mu_1 \neq \mu_2$

Test statistic:

$$t = \frac{\bar{x}_2 - \bar{x}_1}{se(\bar{x}_2 - \bar{x}_1)}$$

We reject H_0 for large values of $|t|$

Under H_0 the test statistic is t-distributed with $n_1 + n_2 - 2$ df

That can be used to compute the (two-sided) P-value: $P = 2\,P(T > |t|)$

where T is t-distributed with $n_1 + n_2 - 2$ df.

In the example we have

$$t = \frac{0.0162}{0.0057} = 2.84$$

corresponding to a P-value of 0.8%

Remark: We have assumed equal variances in the two groups, $\sigma_1^2 = \sigma_2^2$.

This is often not reasonable, in which case a modification of the t-test is appropriate, see for instance section 3.1.9.

In our case, with $n_1 = n_2$, there will be little change, but with very different samples sizes and very different standard deviations in the two groups the unequal variance t-test is recommended.

You may compare the differences between the equal variance and unequal variance t-test in the computer class.

Robustness

All statistical methods are based on some assumptions on the probability model used.

A method is robust if it is valid also when the modeling assumptions do not hold.

The confidence intervals and tests we have considered assume that the observations come from normal distribution(s).

It turns out, however, that the methods are quite robust to the normality assumption when the number of observations is reasonably large.

This is due to the central limit theorem.

Bootstrapping

However, with very small data sets robustness need not hold.

One remedy is “bootstrapping”:

- Resample *with replacement* from original data
- Calculate statistic on this *bootstrap data set*
- Repeat previous 2 steps B (say 1000) times
- Sort the B bootstrap estimates
- Bootstrap (percentile) 95% CI from 2.5 percentile to 97.5 percentile of bootstrap estimates

Very general approach, can be used toward any statistic.

Computationally more demanding.

More and more often implemented in statistical software.

Better CI’s than the percentile interval exist.

Example: Age of minerals

On slide 18 we found 95% CI = (263.8,290.0) using the t-distribution

In comparison, bootstrapping with B=1000 bootstrap samples, I obtained the interval (266.3, 288.9)

The bootstrap intervals are sampled and will differ somewhat when repeated

On slide 3 we found the median of these data to be 273.

Bootstrapping can be also be applied to find a CI for the median.
I obtained the interval (256, 287)