Introduction to statistical inference

Basic Course in Statistic Fall 2007

Thomas Alexander Gerds
tag@biostat.ku.dk

Probability

- in everyday life
- in medical research
- principles and simple rules

Inference

- research questions
- statistical models
- research answers

Probability in everyday life

Example 1: Gambling

Everyone knows that the probability of winning the lottery is a pretty big long shot. How long, however, you probably never really thought about. Your actual odds of winning the lottery depend on where you play, but single state lotteries usually have odds of about 18 million to 1 while multiple state lotteries have odds as high as 120 million to 1.[1]

[1] source: http://www.savingadvice.com/forums/other/5559-probability-winning-lottery-dont-waste-your-money.html

Probability in everyday life

Example 2: Weather

The question I was considering was: what does "70% chance of rain tomorrow" actually mean? Most people would probably expect that if this forecast was issued 100 times, rain would follow about 70 times. And indeed this is what the forecaster thinks (hopes). But on any particular such day, another forecaster might give a different prediction (say "90% chance of rain") and their forecasts might also work out to be accurate on average. Were they both right? What is the "correct" probability of rain?[2]

[2] http://www.thestalwart.com/the_stalwart/2006/01/probability_and_1.html

Medical research

Probability plays an important role

- Prevalence

  How many people have depression?
- Incidence

  How many people develop cancer in the next 10 years?
- Diagnosis – recognizing a disease

  How likely has this person Crohn's disease?
- Prediction – quantifying the risk

  What are the survival chances of a cancer patient?

Example 3: Medical research

Risk Assessment Tool for estimating your 10-year risk of having a heart attack. The risk assessment tool below uses information from the Framingham Heart Study to predict a person's chance of having a heart attack in the next 10 years. This tool is designed for adults aged 20 and older who do not have heart disease or diabetes. To find your risk score, enter your information in the calculator below.

source: http://hp2010.nhlbihin.net/atpiii/calculator.asp

Probability in everyday life

Input to the risk assessment tool

- Age: years
- Gender: Female Male
- Total Cholesterol: mg/dL HDL
- Cholesterol: mg/dL
- Smoker: No Yes
- Systolic Blood Pressure: mm/Hg
- Are you currently on any medication to treat high blood pressure. No Yes

source: http://hp2010.nhlbihin.net/atpiii/calculator.asp

Information about your risk score:

- Age: 30
- Gender: male
- Total Cholesterol: 200 mg/dL
- HDL Cholesterol: 100 mg/dL
- Smoker: Yes
- Systolic Blood Pressure: 90 mm/Hg
- On medication for HBP: Yes

Risk Score* 1%

Means 1 of 100 people with this level of risk will have a heart attack in the next 10 years. * Your risk score was calculated using an equation.

source: http://hp2010.nhlbihin.net/atpiii/calculator.asp

Information about cholesterol:

HDL cholesterol - High density lipoproteins (HDL) is the 'good' cholesterol. HDL carry cholesterol in the blood from other parts of the body back to the liver, which leads to its removal from the body. So HDL help keep cholesterol from building up in the walls of the arteries.

Here are the HDL-Cholesterol Levels that matter to you:

- Less than 40 mg/dL : A major risk factor for heart disease
- 40 to 59 mg/dL : The higher your HDL, the better
- 60 mg/dL and above : An HDL of 60 mg/dL and above is considered protective against heart disease.

source: http://hp2010.nhlbihin.net/atpiii/calculator.asp

Information about your risk score:

- Age: 30
- Gender: male
- Total Cholesterol: 200 mg/dL
- HDL Cholesterol: 40 mg/dL
- Smoker: Yes
- Systolic Blood Pressure: 90 mm/Hg
- On medication for HBP: Yes

Risk Score* 2%

Means 2 of 100 people with this level of risk will have a heart attack in the next 10 years. * Your risk score was calculated using an equation.

source: http://hp2010.nhlbihin.net/atpiii/calculator.asp

What is behind the 'equation'

A logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

- p_i the risk of a heart attack in the next 10 years
- $x_{1,i}$ first factor for subject i: (e.g. age $=$ 30)
- $x_{2,i}$ second factor for subject i: (e.g. gender $=$ male)
- $\ldots$
- $x_{k,i}$ k'th factor for subject i: (e.g. on medication $=$ Yes)
- $\beta_0, \ldots, \beta_k$ regression coefficients estimated based on the Framingham Heart Study data

Next ...

... Simple rules

I $Prob(A \text{ or } B) = Prob(A) + Prob(B)$

II $Prob(A \text{ and } B) = Prob(A) * Prob(B)$
if A and B are independent

III $Prob(A \text{ and } B) = Prob(A \mid B) * Prob(B)$
$= Prob(B \mid A) * Prob(A)$

... Statistical principles

I Relative frequency

II Parameter estimation

III Testing hypothesis

Simple rule I

For a given event, for any two outcomes that might happen the probability of **either** occurring is the **sum** of the individual probabilities.

Example

If

$$Prob(\text{blue eyes}) = 40\%$$

$$Prob(\text{brown eyes}) = 45\%$$

then

$$Prob(\text{brown or blue eyes}) = 85\%$$

Simple rule II

If we consider two or more different events which are **independent** of each other, then to get the probability of a combination of specific outcomes for each of the events we must **multiply** the individual probabilities of those outcomes.

Example

If Mette and Jytte are unrelated and

$$Prob(\text{Mette's first child is a girl}) = 49.9\%$$

$$Prob(\text{Jytte's first child is a girl}) = 49.9\%$$

then

$$Prob(\text{Both are girls}) = 49.9\% * 49.9\% = 24.9\%$$

Simple rule III

For two **dependent** events the probability of a combination of specific outcomes for each of the events we **multiply** the conditional probability of the outcome of the second event **given** the outcome of the first with the unconditional probability of the outcome of the first event.

Example

If

$$Prob(\text{diseased}) = 4\%$$

and

$$Prob(\text{test positive given diseased}) = 81\%$$

then

$$Prob(\text{test positive and diseased}) = 4\% * 81\% = 3.2\%$$

Statistical principles I

Relative frequency

In statistics the frequency of an event is the number of times the event occurred in the experiment or the study. It approximates the probability of the event in the population.

Example: smoking prevalence

- Survey study

$$\frac{\text{no. of daily smokers}}{\text{no. of persons}} \approx \text{Prob(smoking)}.$$

Excursion: Binomial distribution

The binomial distribution is the discrete probability distribution of the number of successes in a sequence of N **independent** yes/no experiments, each of which yields success with probability p

Proportion

$$\hat{p} = \frac{\sum_{i=1}^{N} X_i}{N}$$

Standard error for proportion

$$SE(\hat{p}) = \sqrt{\hat{p} * (1 - \hat{p})/N}$$

Probability to see k cases

$$\binom{N}{k} \hat{p}^k (1 - \hat{p})^{N-k}$$

Confidence interval for proportion

$$\mathrm{CI}_{95\%} = [\hat{p} - 1.96 * SE; \hat{p} + 1.96 * SE]$$

Worked example:

Nr	Daily smoker	x
1	no	0
2	no	0
3	yes	1
4	yes	1
5	no	0
6	yes	1
7	no	0
8	yes	1
9	no	0
10	no	0
11	yes	1
12	yes	1
13	no	0
14	yes	1
15	no	0
16	no	0
17	no	0
18	no	0
19	yes	1
20	no	0
21	no	0
22	no	0

Proportion

$$\hat{p} = \frac{8}{22} = 0.364$$

Probability to see exactly 4 smoker

$$\binom{22}{4} \left(\frac{8}{22}\right)^{k} \left(1 - \frac{8}{22}\right)^{22-4} = 0.037$$

Standard error for proportion

$$SE = \sqrt{\frac{8}{22} * \left(1 - \frac{8}{22}\right) /22} = 0.103$$

Confidence interval for proportion

$$\mathrm{CI}_{95\%} = [0.16; 0.56]$$

Statistical principles II

Parameter[3] estimation

Approximation of an unknown quantity based on data from an experiment or a study.

Example: linear regression

$$\text{Systolic blood pressure} = \beta_0 + \beta_1 * \text{male} + \text{unexplained variation}$$

The regression coefficients (β_0, β_1) are unknown parameters.

[3] Parameters are quantities that define certain relatively constant characteristics of systems or functions.

Worked example: systolic blood pressure

Systolic blood pressure

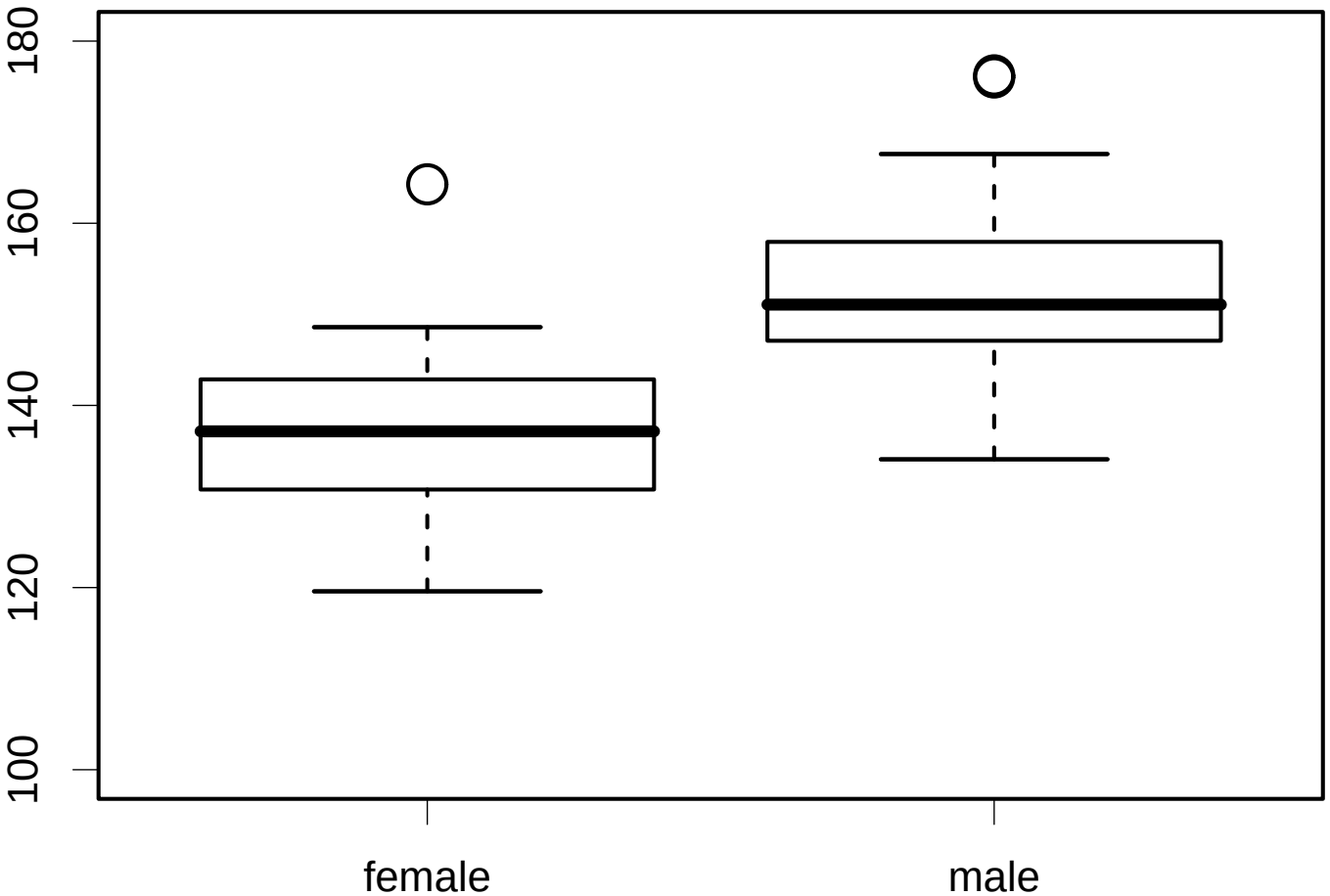
- male (n=24)

  145.8,142.4,156.0,150.9,146.1,145.4,139.3,154.8,
  129.5,146.5,144.3,153.5,166.8,148.4,131.3,167.4,
  131.4,141.9,154.4,145.6,141.7,144.3,130.3,165.4

- female (n=29)

  151.8,130.0,156.8,126.4,142.0,145.2,118.7,144.8,
  150.4,149.9,129.9,127.4,137.3,143.9,142.0,133.4,
  164.5,139.0,153.6,141.3,134.7,119.4,130.6,152.7,
  110.2,150.2,145.6,138.5,145.0

Display the data



Linear regression model

Estimates are the group-wise mean systolic blood pressure values

$$\hat{\beta}_0 = \frac{1}{n} \sum_{i:females} X_i = \frac{1}{29} \sum_{i=1}^{29} X_i = 139.83$$

$$\hat{\beta}_1 = \frac{1}{n} \sum_{i:males} X_i - \hat{\beta}_0 = \frac{1}{24} \sum_{i=1}^{24} X_i - 139.83 = 146.81 - 139.83 = 6.98$$

How reliable are these estimates?

Statistical uncertainty

The variability of the parameter estimate depends on

- the sample size
- the statistic/estimate (here we used the mean)
- the variability of the measurements in the population
- the measurement error
- how much of the variability in the data can be explained by other measured factors

Displaying the statistical uncertainty

The sample **standard deviation**

$$SD = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \overline{X})^2}; \qquad \overline{X} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

measures the variability of the **measurements** (the sample variance is SD^2).

The estimated **standard error** of the mean

$$SE = SD/\sqrt{N}$$

measures the variability of the **mean** around the population value.

Reporting the statistical uncertainty

A **confidence interval** is a range of values which we can be confident includes the true value.

Most common construction

A **95% confidence interval** for the parameter β is

$$[\hat{\beta} - z_{1-\alpha/2} * SE; \hat{\beta} + z_{1-\alpha/2} * SE]$$

where $\hat{\beta}$ is the parameter estimate, SE is an estimate of its standard error, $\alpha = 0.05$ and $z_{1-\alpha/2}$ depends on the sampling distribution of $\hat{\beta}$.

Worked example

Parameter estimate

$$\hat{\beta}_1 = 6.98$$

Standard error

$$SE = 3.21$$

For $\alpha = 5\%$

$z_{1-\alpha/2} = 2.01$ t-distribution with 51 degrees of freedom

$(z_{1-\alpha/2} = 1.96$ normal distribution$)$

A 95% confidence interval for the effect of gender on the systolic blood pressure is given by

$$\mathrm{CI}_{95\%} = [0.52; 13.4]$$

Statistical principles III

Testing a hypothesis[4]

Accept or reject a hypothesis based on data.

Example:

- *Null hypothesis*: The systolic blood pressure is the same for females and males
- *Alternative hypothesis*: The systolic blood pressure is different for females and males

[4] A hypothesis consists either of a suggested explanation for a phenomenon or of a reasoned proposal suggesting a possible correlation between multiple phenomena.

Excursion

The same hypothesis in the linear regression model

$$\text{Systolic blood pressure} = \beta_0 + \beta_1 * \text{male} + \text{unexplained variation}$$

- *Null hypothesis*: $\beta_1 = 0$
- *Alternative hypothesis*: $\beta_1 \neq 0$

Testing the hypothesis

Students t-test statistic[5]

$$T = \frac{|\ mean(group\ 1)\text{-}mean(group\ 2)\ |}{standard\ error}$$

Decision: if T is large reject the hypothesis, if T is small accept the hypothesis

Equivalently: if the p-value is small reject the hypothesis, if the p-value is large accept the hypothesis

[5] The statistic was introduced by William Sealy Gosset for cheaply monitoring the quality of beer brews. "Student" was his pen name.

Reporting the result of a statistical test

The **p-value** is the probability of having observed the data (or more extreme data) **when the null hypothesis is true**. The result of the test is called statistically significant at the significance level $\alpha = 5\%$ if the p-value is smaller than 0.05.

Worked example

The t-test statistic is

$$T = \frac{146.81 - 139.83}{3.22} = 2.17$$

The p-value is

$$p = 0.0347$$

Since $p < 0.05$ there is a significant gender effect on the systolic blood pressure (at the usual 5% significance level).

Relation between confidence intervals and p-values

If we estimate a real value β and have computed a 95% confidence interval $[a, b]$ then the null hypothesis

$$\beta = 0$$

can be rejected at the 5% significance level if 0 is **not** in $[a, b]$.

- 5% and 95% can be replaced by α and $1 - \alpha$
- 0 can be replaced by any other possible value of the parameter β.

Statistical inference. . .

. . . starts with a medical research question in a defined population. . .

. . . finds an appropriate statistical model and samples enough data. . .

. . . draws conclusions based on data analysis. Unfortunately these conclusions include statistical certainty.

Finding medical research

Leading medical journals

- The New England Journal of Medicine
- Journal of the American Medical Association
- Nature Medicine
- The Lancet

Journals in related fields

- Statistics in medicine
- American Journal of Epidemiology
- Bioinformatics

The Lancet, Current Issue, Volume 370, Number 9594, 6 October 2007

Zinc supplementation does not significantly affect child mortality in Nepal
Summary — Full Text

Thalidomide improves survival of elderly with multiple myeloma when added to standard chemotherapy regimen
Summary — Full Text

Daytime and night-time blood pressure are both vital prognostic indicators
Summary — Full Text

The summary of a typical medical article contains

1. background/objective
2. methods/resources
3. results/findings
4. conclusions/interpretation

A good research question considers all these points from the very beginning. You should improve your own research question by consulting the literature, experts in your and related fields, and a statistician.

Example 1: The effectiveness of supported employment for people with severe mental illness: a randomised controlled trial.[6]

BACKGROUND: The value of the individual placement and support (IPS) programme in helping people with severe mental illness gain open employment is unknown in Europe. Our aim was to assess the effectiveness of IPS, and to examine whether its effect is modified by local labour markets and welfare systems. METHODS: 312 patients with severe mental illness were randomly assigned in six European centres to receive IPS (n=156) or vocational services (n=156). Patients were followed up for 18 months. The primary outcome was the difference between the proportions of people entering competitive employment in the two groups. The heterogeneity of IPS effectiveness was explored with prospective meta-analyses to establish the effect of local welfare systems and labour markets. Analysis was by intention to treat. (...) FINDINGS: IPS was more effective than vocational services for every vocational outcome, with 85 (55%) patients assigned to IPS working for at least 1 day compared with 43 (28%) patients assigned to vocational services (difference 26.9%, 95% CI 16.4-37.4). Patients assigned to vocational services were significantly more likely to drop out of the service and to be readmitted to hospital than were those assigned to IPS (...) INTERPRETATION: Our demonstration of the effectiveness of IPS (...) confirms this service to be an effective approach for vocational rehabilitation in mental health that deserves investment (...)

[6] Burns et al., Lancet. 2007 Sep 29;370(9593):1108-9.

Example 2: Very low birth weight increases risk for sleep-disordered breathing in young adulthood: the Helsinki Study of Very Low Birth Weight Adults. [7]

OBJECTIVE: We investigated whether very low birth weight ($<$ 1500 g) is associated with the risk of sleep-disordered breathing in young adulthood. METHODS: The study was a retrospective longitudinal study of 158 young adults born with very low birth weight and 169 term-born control subjects (aged 18.5-27.1 years). The principal outcome variable was sleep-disordered breathing defined as chronic snoring. RESULTS: The crude prevalence of chronic snoring was similar in both groups: 15.8% for the very low birth weight group versus 13.6% for the control group. However, after controlling for the confounding variables in multivariate logistic regression models (age, gender, current smoking, parental education, height, BMI, and depression), chronic snoring was 2.2 times more likely in the very low birth weight group compared with the control group. In addition, maternal smoking during pregnancy was significantly and independently of very low birth weight related to risk of sleep-disordered breathing. Maternal preeclampsia, standardized birth weight, and, for very low birth weight infants, small-for-gestational-age status were not related to sleep-disordered breathing. CONCLUSIONS: Premature infants with very low birth weight have a twofold risk of sleep-disordered breathing as young adults. In addition, maternal smoking during pregnancy increases the risk of sleep-disordered breathing by more than twofold.

[7] Paavonen et al., Pediatrics. 2007 Oct;120(4):778-84.

The statistical model...

...formulates parameters and hypotheses...

...in terms of the **probability distribution** of the data...

...so that the medical research question would be answered if the parameters and hypotheses were known.

Normal distribution model for a continuous variable

Research question: what is the probability that a man has systolic blood pressure higher than 140?

Statistical model:

- Parameter

$$F(x) = Prob(systolic\ blood\ pressure \leq x)$$

  F is the cumulative distribution function.

- Assumption
  The systolic blood pressure in the male population follows a normal distribution with unknown mean μ and variance σ^2

$$\mathcal{N}(\mu, \sigma).$$

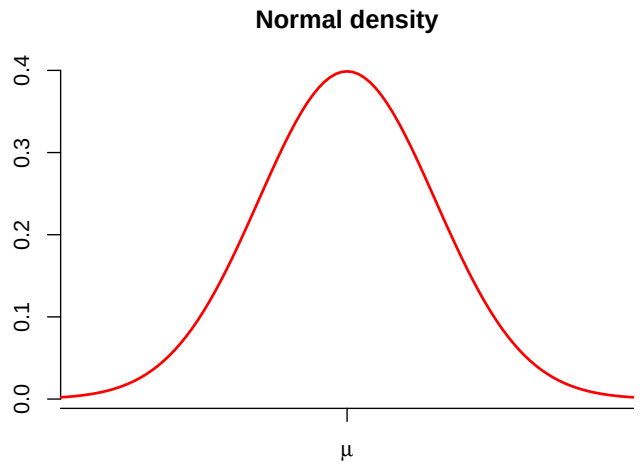
The normal distribution

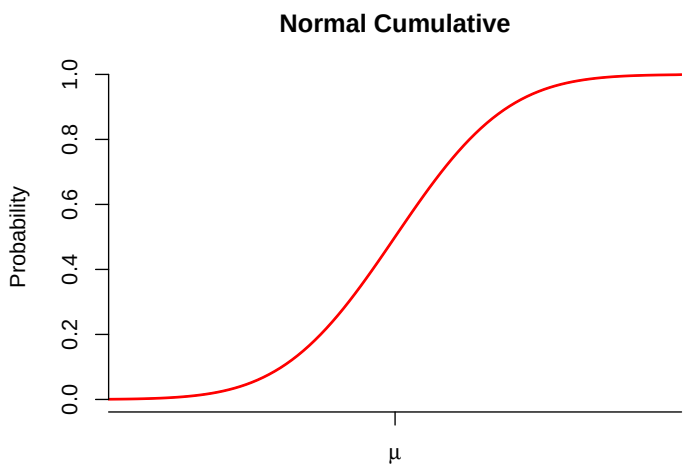
has density

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

and cumulative distribution function

$$F(x) = \int_{-\infty}^{x} \varphi_{\mu,\sigma^2}(u)\mathrm{d}u.$$





Worked example

The mean of the systolic blood pressure of the 24 men in our example is $\hat{\mu} = 146.81$ the standard deviation in this group is $SD = \hat{\sigma} = 10.70$ Thus in the normal model the answer is

$$F(140) = \int_{-\infty}^{140} \frac{1}{\hat{\sigma}\sqrt{2\pi}} \exp\left(-\frac{(u-\hat{\mu})^2}{2\hat{\sigma}^2}\right) \mathrm{d}u$$

$$= \int_{-\infty}^{140} \frac{1}{10.7\sqrt{2\pi}} \exp\left(-\frac{(u-146.81)^2}{2 * 10.7^2}\right) \mathrm{d}u$$

$$= 26\%.$$

Value of the result?

More complex research questions

- Does the blood pressure depend on gender, age, medication, . . . ?
- Does the risk of chronic snoring depend on the birth weight, parental education, . . . ?

Regression models

These questions require a model for the conditional distribution of a dependent variable given factors

- Linear regression model

$$Y_i = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

- Logistic regression model

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_k x_{k,i}.$$

Summary

Medical side	Statistical side
Research question	Model, parameters, hypotheses
Examinations, data collection	Data cleaning, data analysis
Clinical significance, interpretation	Descriptive statistics, p-values, confidence intervals