NMAI059 Probability and statistics 1
Class 13

Robert Šámal

Statistics – What have we learnt

- basic setup: we consider random sample $X_1, \ldots, X_n$ from distribution F_ϑ — describes the measurement process, all ways how could it go
- we measure data – particular numbers $x_1, \ldots, x_n$, so called realization of random sample, — what did we really measure

1. point estimation: determine best possible number, estimate for the parameter ϑ, or some function of it, $g(\vartheta)$.
2. interval estimation: determine an interval, that contains the unknown parameter ϑ with a large probability
3. hypothesis testing

Overview

Hypothesis testing

Goodness of fit test

Linear regression

Hypothesis testing – illustration

- We want to test, if a coin is fair.
- H_0: it is fair
- H_1: not fair ("Scientists discovered, that casino XY uses loaded coin.")
- Results: Reject H_0/don't reject H_0
- Type I error: false rejection. We reject H_0, even if it is true. Embarassing.
- Type II error: false non-rejection. We don't reject H_0, even if it is false. Unused opportunity.
- Need to find k such that we will reject H_0 if $|S - n/2| > k$.

Hypothesis testing – general approach

- We choose an appropriate statistical model.
- We choose *significance level* α: prob. of false rejection of H_0. Typically $\alpha = 0.05$ (medicine/psychology – much less in high-energhy physics).
- We determine *test statistics* $S = h(X_1, \dots, X_n)$, that we will determine from the measured data.
- We determine *rejection region* – set W.
- We measure $x_1, \dots, x_n$ – so-called realizations of $X_1, \dots, X_n$.
- Decision rule: we reject H_0 iff $h(x_1, \dots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$
- $\beta = P(h(X) \notin W; H_1)$ … *strength of the test*

- often we do not choose α in advance but compute so-called p-*value:* minimal α, for which we would reject H_0.

Hypothesis testing – an example

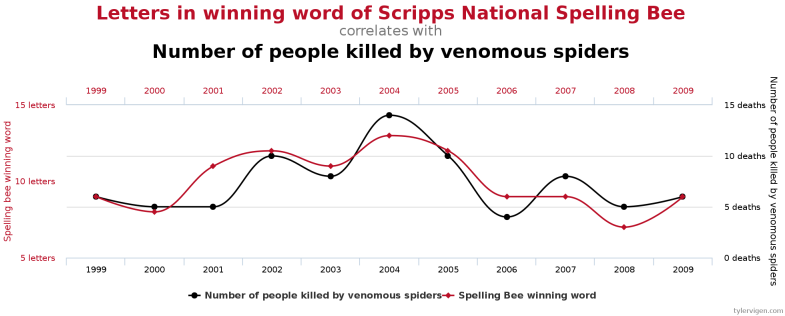
- $X_1, \ldots, X_n$ random sample from $N(\vartheta, \sigma^2)$
- σ^2 known
- $H_0 : \vartheta = 0 \qquad H_1 : \vartheta \neq 0$

Hypothesis testing – an example

- $X_1, \ldots, X_{n_1}$ random sample from $Ber(\vartheta_X)$
- $Y_1, \ldots, Y_{n_2}$ random sample from $Ber(\vartheta_Y)$
- $H_0 : \vartheta_X = \vartheta_Y \qquad H_1 : \vartheta_X \neq \vartheta_Y$

p-hacking

- we first gain data, then look for interesting stuff
- – given enough data, there will be random coincidences
- even worse, we may test, until we get the desired outcome
- *reproducibility* – after exploratory analysis of the data we make an independent measurement and a confirmatory analysis.
- or we split the data in advance to a part for hypothesis formation and part for verification . . . simple example of cross validation



Overview

Hypothesis testing

Goodness of fit test

Linear regression

χ^2_k – chi-square distribution

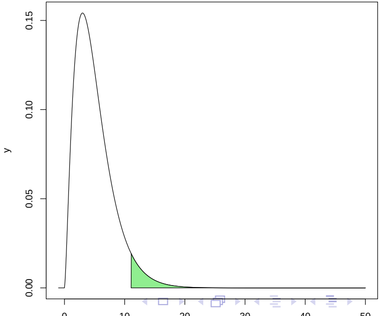
Definition

$Z_1, \ldots, Z_k \sim N(0,1)$ i.i.d. The distribution of r.v.

$$Q = Z_1^2 + \cdots + Z_k^2$$

is called chi-square(d) with k degrees of freedom (really k!), and denoted χ^2_k.

- $\mathbb{E}(Q) = k$ (easy)
- $var(Q) = 2k$ (fyi, you don't have to remember this)
- density can be written by a reasonable formula
- $Q \doteq N(k, 2k)$ for large k (CLT)



Multinomial and categorical distribution

Definition

Given $p_1, \ldots, p_k \geq 0$ *so, that* $p_1 + p_2 + \cdots + p_k = 1$*.*
we repeat n*-times an experiment with* k *possible outcomes, where the* i*th has probability* p_i
$X_i :=$ *how many times we got the* i*-th outcome* $(X_1, \ldots, X_k)$ *has multinomial distribution with parameters* $n, (p_1, \ldots, p_k)$*.*

- trivial example: $X_i =$ number of die rolls that equaled i
- important example: $X_i =$ number of occurences of i-th letter,
- $P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}$

Pearson χ^2 statistics

- $(X_1, \ldots, X_k)$ – multinomial distribution with parameters $n, (p_1, \ldots, p_k)$ as above
- $E_i := \mathbb{E}\big(X_i\big) = np_i$
- *Pearson* χ^2 *statistics* is the function

$$\chi^2 = T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$$

- **Theorem** $T \xrightarrow{d} \chi^2_{k-1}$

Goodness of fit test

- $(X_1, \ldots, X_k)$ – multinomial distribution with parameters $n, (\vartheta_1, \ldots, \vartheta_k)$ as above
- n known, ϑ unknown
- Null hypothesis H_0: $\vartheta = \vartheta^*$ for some given ϑ^*
- $E_i := n\vartheta_i^*$ for all i
- We use the statistics $\chi^2 = T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$
- We reject H_0 iff $T > \gamma$
- $\gamma := F_Q^{-1}(1-\alpha)$, where $Q \sim \chi^2_{k-1}$
- $P($I type error$) = P(T > \gamma; H_0) \to P(Q > \gamma) = \alpha$

Goodness of fit test – example

- We roll a die repeatedly (600 times). The numbers 1 upto 6 came up with frequencies 92, 120, 88, 98, 95, 107.
- Is the die fair?

Extensions

- To study a distribution of an arbitrary r.v. Y we can pick "bins" $B_1, \ldots, B_k$ (a partition of $\mathbb{R}$) and look how often $Y \in B_i$ (this will be measured by r.v. X_i).
- Similar test for independence of discrete random variables.

Overview

Hypothesis testing

Goodness of fit test

Linear regression

Linear regression – the problem

- data: (x_i, y_i) for $i = 1, \dots, n$
- goal: $y = \vartheta_0 + \vartheta_1 x$

- we measure how good fit we have by the quadratic error:

$$\sum_{i=1}^{n} \big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$

Linear regression – solution

- To minimalize
$$\sum_{i=1}^{n}\big(y_i - (\vartheta_0 + \vartheta_1 x_i)\big)^2$$
- the optimal parameters are
$$\hat{\vartheta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \qquad \hat{\vartheta}_0 = \bar{y} - \vartheta_1 \bar{x},$$

  where $\bar{x} := (x_1 + \cdots + x_n)/n$, $\bar{y} := (y_1 + \cdots + y_n)/n$.

Linear regression – why sum of squares?

- We assume that $x_1, \dots, x_n$ are fixed, y_i is a realization of a r.v.

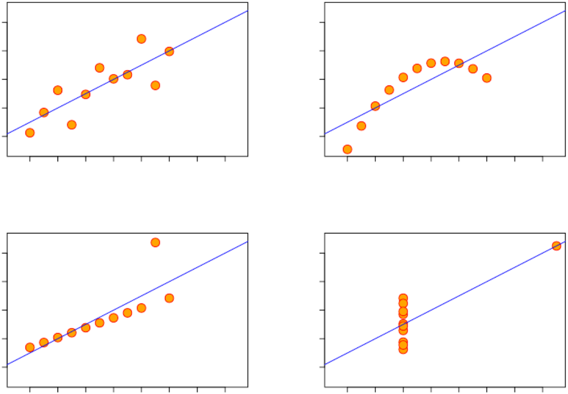
$$Y_i = \vartheta_0 + \vartheta_1 x_i + W_i$$

- $W_i \sim N(0, \sigma^2)$ for all i; $W_1, \dots, W_k$ iid
- maximal likelihood:

$$L(y;\vartheta) = \prod_{i=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y_i-\vartheta_0-\vartheta_1 x_i)^2}{2\sigma^2}}$$

- $\ell(y;\vartheta) = \log L(y;\vartheta) = a + b\sum_{i=1}^{n}(y_i - \vartheta_0 - \vartheta_1 x_i)^2$

Limits of regression



(data: Francis Anscombe 1973, image: wikieditor Schutz)

- nonlinear regression
- logistic regression

Simpson's paradox

Stone size \ Treatment	Treatment A	Treatment B
Small stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)
Both	78% (273/350)	**83% (289/350)**