NMAI059 Probability and statistics 1
Class 13

Robert Šámal

Statistics – What have we learnt

- basic setup: we consider random sample $\overbrace{X_1, \dots, X_n}^{\text{i.i.d}}$ from distribution F_ϑ — describes the measurement process, all ways how could it go
- we measure data – particular numbers $x_1, \dots, x_n$, 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

but fixed

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 $|S - n/2| > k$.

toss coin n-times $n = 1000$ $S := \#$ of heads

$S = 200$ $p \doteq 0$

Hypothesis testing – general approach

$X_1, \dots, X_n \sim Ber(p)$, iid
$X = X_1 + \dots + X_n \sim Bin(n, p)$

- 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-energly physics).
- We determine *test statistics* $S = h(X_1, \dots, X_n)$, that we will determine from the measured data.
  $= \sum X_i \sim Bin(n,p)$
- We determine *rejection region* – set W.
  $W = \{x : |\frac{x}{n} - \frac{1}{2}| > k\}$
- We measure $x_1, \dots, x_n$ – so-called realizations of $X_1, \dots, X_n$.
- Decision rule: we reject H_0 iff $h(x_1, \dots, x_n) \in W$.
- $\alpha = P(h(X) \in W; H_0)$ P(Type I error)
- $\beta = P(h(X) \notin W; H_1)$ ... *strength of the test* $1 - \beta$
  P(Type II error)
- 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

measure temp.

- $X_1, \ldots, X_n$ random sample from $N(\vartheta, \sigma^2)$

mean = true temp.

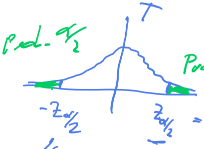
- σ^2 known
- $H_0 : \vartheta = 0 \qquad H_1 : \vartheta \neq 0$

$\alpha = 0.05$

Assume $H_0 : \vartheta = 0$

$Z := \overline{X}_n \sim N(\vartheta, \sigma^2/n)$

$T := \dfrac{\overline{X}_n}{\sigma/\sqrt{n}} \sim N(0,1)$ .... the distr. of test stat., assume H_0, should be known (at least approx.)



$z_{\alpha/2} = \Phi^{-1}(1-\frac{\alpha}{2})$

$-z_{\alpha/2} = \Phi^{-1}(\alpha/2)$

$\boxed{W = (-\infty, -z_{\alpha/2}) \cup (z_{\alpha/2}, +\infty)}$

If $T \in W$, we reject H_0

Goal: $P(T \in W; H_0) = \alpha$

Hypothesis testing – an example

two-sample test

ind.

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

$X_i = [i\text{-th male patient recovers}]$

$Y_j = [j\text{-th female recovered}]$

$\hat\Theta_X = \bar X_{n_1} = \frac{X_1 + \dots + X_{n_1}}{n_1} \qquad \hat\Theta_Y = \bar Y_{n_2} = \frac{Y_1 + \dots + Y_{n_2}}{n_2}$ ... approx. normal by CLT

we estimate $\vartheta_X - \vartheta_Y = \hat\Theta_X - \hat\Theta_Y =: Z$ if this is too large, we reject H_0

$EZ = E(\hat\Theta_X) - E(\hat\Theta_Y) = \vartheta_X - \vartheta_Y = 0$ ASSUMING H_0 $\qquad \vartheta = \vartheta_X = \vartheta_Y$

$var(Z) = var(\hat\Theta_X - \hat\Theta_Y) \overset{ind.}{=} var(\hat\Theta_X) + var(\hat\Theta_Y) = \frac{var(X_1)\cdot n_1}{n_1^2} + \frac{var(Y_1)\cdot n_2}{n_2^2}$

$= \sigma^2$

$= \frac{\vartheta_X \cdot (1-\vartheta_X)}{n_1} + \frac{\vartheta_Y (1-\vartheta_Y)}{n_2} = \vartheta(1-\vartheta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

we estimate ϑ by $\hat\Theta = \frac{\sum X_i + \sum Y_j}{n_1 + n_2}$

$\hat\sigma^2 = \hat\Theta(1-\hat\Theta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

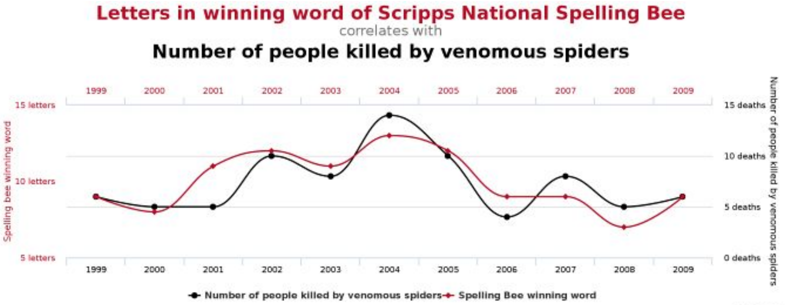
$T = \frac{\hat\Theta_X - \hat\Theta_Y}{\sqrt{\hat\Theta(1-\hat\Theta)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{approx}{\sim} N(0,1)$

$\boxed{P(T \in W) \doteq \alpha}$

$\hat\Theta_X, \hat\Theta_Y$ – approx. normal $\longrightarrow$ $\hat\Theta_X - \hat\Theta_Y$ approx. normal

p-hacking BAD

- 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

χ_k^2 – chi-square distribution

$E(Q) = E(Z_1^2) + \dots + E(Z_k^2)$ (linear)
$= 1 + \dots + 1 = k$

Definition

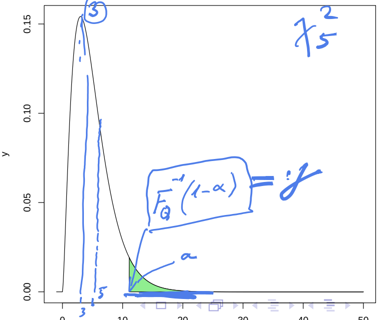
$Z_1, \dots, Z_k \sim N(0,1)$ i.i.d. The distribution of r.v.

$E(Z_i) = 0$
$var(Z_i) = 1 = E(Z_i^2) - (E Z_i)^2$, $E Z_i = 0$

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

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

- $\mathbb{E}(Q) = k$ (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

experim. with k poss. outcomes

type 1
type 2
⋮
type k

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,
- $P(X_1 = x_1, \ldots, X_k = x_k) = \binom{n}{x_1, \ldots, x_k} p_1^{x_1} \ldots p_k^{x_k}$

gdy $x_1 + \ldots + x_k = n$

$X_1 + \ldots + X_k = n$

Pearson χ^2 statistics

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

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

$T_n = (Z_n)^2$ $\quad Z_n \xrightarrow{d} N(0,1)$

$Z_n^2 \xrightarrow{d} N(0,1)^2 = \chi^2_1$

k fixed; as $n \to \infty$

O_i -- observed

$X_1 \sim Bin(n, p_1)$

$\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}} \overset{approx.}{\sim} N(0,1)$ CLT

(the limit does not dep. on $p_1 \dots p_k$)

$np_1 - X_1$

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

sketch of proof for $k=2$

$X_1 + X_2 = n$

$p_1 + p_2 = 1$

$E_i = np_i$

$$T = \frac{(X_1 - np_1)^2}{np_1} + \frac{(X_2 - np_2)^2}{np_2} = \frac{(X_1 - np_1)^2}{np_1} + \frac{(\overbrace{n - X_1 - n(1-p_1)}^{np_1 - X_1})^2}{np_2}$$

$$\frac{(X_1 - np_1)^2}{n} \cdot \left(\frac{1}{p_1} + \frac{1}{p_2}\right) = \frac{(X_1 - np_1)^2}{n p_1 p_2} = \left(\frac{X_1 - np_1}{\sqrt{np_1(1-p_1)}}\right)^2$$

Goodness of fit test

- $(X_1, \dots, X_k)$ – multinomial distribution with parameters $n, (\vartheta_1, \dots, \vartheta_k)$ 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 $\chi^2 = T := \sum_{i=1}^{k} \frac{(X_i - E_i)^2}{E_i}$

  approx.: $T \sim \chi^2_{k-1}$
- We reject H_0 iff $T > \gamma$
- $\boxed{\gamma := F_Q^{-1}(1-\alpha)}$, where $Q \sim \chi^2_{k-1}$
- $P(\text{I type error}) = P(T > \gamma; H_0) \to P(Q > \gamma) = \alpha$

  $n \to \infty$

  $1 - F_T(\gamma) \to 1 - F_Q(\gamma)$

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.

  $\underset{x_1}{\overset{4}{92}}$, $\underset{x_2}{\overset{4}{120}}$, ... x_6

- Is the die fair?

$n = 600 \quad \vartheta^* = \left(\frac{1}{6}, \dots, \frac{1}{6}\right)$

$E_i = n\vartheta^* = 100$

$$T = \sum_{i=1}^{6} \frac{(X_i - E_i)^2}{E_i} = \sum \frac{(X_i - 100)^2}{100} = \frac{8^2}{100} + \frac{20^2}{100} + \frac{12^2}{100} + \frac{2^2}{100} + \frac{5^2}{100} + \frac{7^2}{100}$$

$$= \underline{6.86} < 11.1 \Rightarrow \text{do not reject } H_0$$

(we trust die is fair)

$Q \sim \chi^2_5$

$F_Q^{-1}(0.95) = 11.1$

$1 - F_Q(6.86)$

p-value $= 0.23$

Extensions

$X_i = [Y \in B_i]$

- 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.

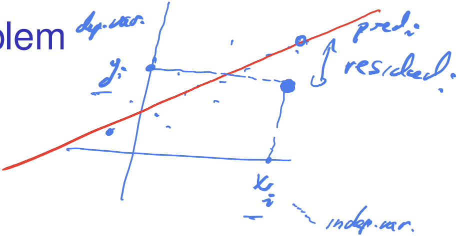
Overview

Hypothesis testing

Goodness of fit test

Linear regression

Linear regression – the problem



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

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

$$\sum_{i=1}^{n} \underbrace{\Big(y_i - \underbrace{(\vartheta_0 + \vartheta_1 x_i)}_{\text{pred.}}\Big)}_{\text{residual.}}{}^2$$

Linear regression – solution

- To minimalize

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

we know from L.A.

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

$\frac{1}{n-1}\sum(x_i - \bar x)(y_i - \bar y)$ — sample covariance

$\frac{1}{n-1}\sum(x_i - \bar x)^2$ — sample variance

Linear regression – why sum of squares?

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

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

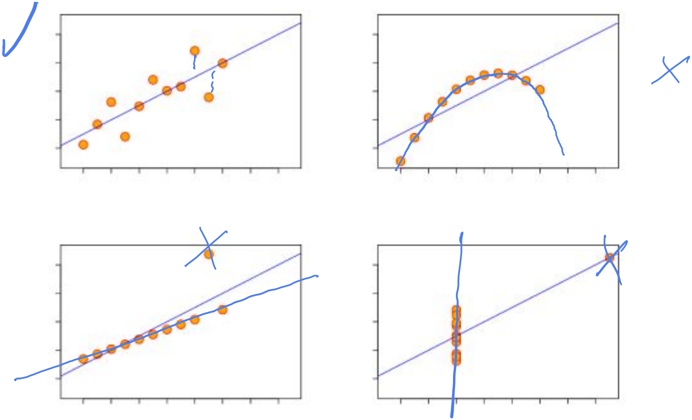
to set the pred. by physical eq.

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

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

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

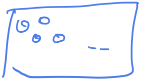
Limits of regression



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

- nonlinear regression
- logistic regression

box w. balls $1, \dots, N$ $N = ?$



$\longrightarrow$ estimate N

$\longrightarrow$ pick 5 balls $x_1, \dots, x_5$ with repet.

moment method: 1.st moment $= \frac{N+1}{2} = \frac{x_1 + \dots + x_5}{5}$

$$\Rightarrow \hat{N} = 2\,\frac{x_1 + \dots + x_5}{5} - 1$$

$$E(\hat{N}) = N$$

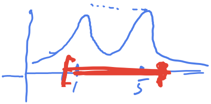
MSE larger than other methods

$1,1,1,1,11 \longrightarrow 3$, but $N \geq 11$

$\hat{N} = 5$

max. likelihood ......

$$\hat{N} = \max\{x_1, \dots, x_5\}$$



$$\hat{N} = \max(x_1 \dots x_5) \cdot \text{const}$$

Simpson’s paradox

Treatment **Stone size**	**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)**