Introduction to statistical concepts

Oliver Muthmann

07/27/2016

Contents

1 Introduction

Problem:

direct — If you have a model, you can generate data .

inverse — If you have data, it is difficult to find the corresponding model.

$$\text{Model} \Rightarrow \text{Data}$$

but

$$\text{Data} \not\Rightarrow \text{Model}.$$

'Solution': go back and forth between direct and inverse problem.

1.1 Model classification

- continous, discrete
- deterministic, stochastic
- linear, nonlinear

1.2 General remarks

- need to understand few things properly
- need to recognize a lot
- need to be able to find the rest in literature

2 Distributions

2.1 Random variables

Random variable X:

- probability density $p_X(x)$
- realization x in $(x, x+dx)$ has probability $p_X(x)dx$.
- $p_X(x) \geq 0, \quad \int p_X(x)dx = 1$

2.2 Moments

$\mu_k = \langle x^k \rangle = \int x^k p(x)dx$

Mean $\mu_1 = \bar{x} = \mu = \langle x \rangle$

Variance $\sigma^2 = \langle (x-\bar{x})^2 \rangle = \mu_2 - \mu_1^2$ standard deviation σ

Skewness $\kappa = \langle (x-\bar{x})^3 \rangle$ measure for asymmetry

Kurtosis $\gamma = \langle (x-\bar{x})^4 \rangle / \sigma^4 - 3$

2.3 Examples of distributions

Gaussian (or normal) distribution $N(\mu, \sigma^2)$:

$$p_G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\bar{x})^2}{2\sigma^2}}$$

Standard normal distribution: $N(0, 1)$
68% of its probability mass in $[-\sigma, \sigma]$
95% of its probability mass in $[-2\sigma, 2\sigma]$
Moments:$\langle x^k \rangle = \begin{cases} 0 & k \text{ even} \\ 1 \cdot 3 ... (k-1) & k \text{ odd} \end{cases}$
Central limit theorem: Sum of independent random variables with finite moments converges to a gaussian distribution.

Uniform distribution $U(a, b)$:

$$p(x) = \begin{cases} 1/(b-a) & a \leq x \leq b \\ 0 & \text{else.} \end{cases}$$

Exponential distribution $Exp(\tau)$:

$$p(x) = \frac{1}{\tau} e^{-x/\tau}$$

Mean and standard deviation:$\mu = \sigma = \tau$
Can use uniform numbers to generate

$$x \sim U(0, 1) \quad \Longrightarrow \quad -\log(x) \sim Exp(1)$$

χ_r^2 distribution with r degrees of freedom:

$$\chi_{r^2}(x) = \sum_{i=1}^{r} (N(0, 1))^2$$

Sum of r independent, standard normal distributed random variables.
Can turn this around to reject the hypothesis that variables were independent.
Mean $\mu = r$
Variance $\sigma^2 = 2r$

t-distribution:

$$t(r, x) = \frac{N(0,1)}{\sqrt{\chi_r^2/r}}$$

To test whether the mean of a normal distribution is different from zero. The test is done by summing up squared differences from zero. That explains the χ_r^2-distribution.

F-distribution:

$$F(r_1, r_2, x) = \frac{\chi_{r_1}^2/r_1}{\chi_{r_2}^2/r_2}$$

To test the variances of two normal distributions. For that, divide the variances and see whether it is roughly one. To obtain variances, we need to sum up squared differences from the mean. Subtract the mean, then we again obtain χ_r^2-distributed numbers.

Cauchy distributionCauchy(x, μ, γ):

$$\text{t(1,x)=Cauchy (x,0,1)=}\frac{N(0,1)}{N(0,1)}$$

Moments do not exist.
μ is a location parameter, not the mean.
Also known as Breit-Wigner or Lorenz distribution.

Binomial distribution:

$$B(n, p, k) = \binom{n}{k} p^k (1-p)^{n-k}, \qquad k \in \{0, 1, ..., n\}$$

Poisson distribution:

$$P(k, \lambda) = \frac{e^{-\lambda}\lambda^k}{k!}$$

Limiting case of the Binomial distribution, with $\lim_{n\to\infty} np = \lambda$

2.4 Estimators

Mean of a Gaussian random variable

Have N realizations x_i from $N(\mu, \sigma^2)$. The *estimator*

$$\hat{\mu} = \frac{1}{N}\sum_{i-1}^{N} x_i$$

is a Gaussian distributed random variable with mean

$$\langle \hat{\mu} \rangle = \mu$$

this estimator is *unbiased* as, on average, it gives the true mean.

$$\mathrm{Var}(\hat{\mu}) = \langle (\hat{\mu} - \langle \hat{\mu} \rangle)^2 \rangle = \frac{1}{N}\sigma^2$$

Confidence interval: with a probability of 95 %, the true value of μ is in the interval

$$\left[\hat{\mu} - 1.96\sqrt{\frac{1}{N}\sigma^2}, \hat{\mu} + 1.96\sqrt{\frac{1}{N}\sigma^2}\right]$$

The confidence interval shortens as $\sqrt{\frac{1}{N}}$. The estimator $\hat{\mu}$ is *consistent.*

Variance of a Gaussian random variable

The estimator

$$\hat{\sigma^2} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \hat{\mu})^2$$

is a χ^2_{N-1} distributed random variable. Only $N-1$ degrees of freedom as we had to estimate μ.