Introduction to statistical concepts

Oliver Muthmann

07/27/2016

Contents

1 Introduction **1**
1.1 Model classification 2
1.2 General remarks 2

2 Distributions **2**
2.1 Random variables 2
2.2 Moments 3
2.3 Examples of distributions 3
2.4 Estimators 5

3 Error propagation, correlations and linear regression **6**
3.1 Absolute error σ_x 6
3.2 Relative error $\frac{\sigma_x}{x}$ 6
3.3 General approximation 6
3.4 Weighted average 7
3.5 Linear regression 7

4 Parametric tests **8**

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:$\langle x^k \rangle = \begin{cases} 0 & k \text{ even} \\ 1 \cdot 3 \ldots (k-1) & k \text{ odd} \end{cases}$
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^2_r 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:

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

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^2_r-distribution.

F-distribution:

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

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^2_r-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.

3 Error propagation, correlations and linear regression

Here, we assume that errors are independent and Gaussian distributed, and we know the variance (given by the specifications of the measurement device or determined in a separate experiment). Want to know what happend to the error when we do computations with the measured quantity.

$$x \to y = f(x) \qquad \sigma_x \to \sigma_y = ?$$

3.1 Absolute error σ_x

- invariant under addition/subtraction of a constant:

$$f(x) = y = x \pm c \quad \Rightarrow \quad \sigma_y = \sigma_x$$

- variance additive under addition/subtraction of two measurements:

$$f(x_1, x_2) = y = x_1 \pm x_2 \quad \Rightarrow \quad \sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2$$

- addition/subtraction of correlated measurements

$$f(x_1, x_2) = y = x_1 \pm x_2 \quad \Rightarrow \quad \sigma_y^2 = \sigma_{x_1}^2 + \sigma_{x_2}^2 \pm 2\mathrm{Cov}(x_1, x_2)$$

3.2 Relative error $\frac{\sigma_x}{x}$

- invariant under multiplication of a constant:

$$f(x) = y = c \cdot x \quad \Rightarrow \quad \frac{\sigma_y}{y} = \frac{\sigma_x}{x}$$

3.3 General approximation

Assumes that errors are small enough that the function can be linearized.

- one measured quantity x

$$y = f(x) \quad \Rightarrow \quad \sigma_y \approx \frac{df}{dx}\sigma_x$$

- two measured quantities x_1, x_2

$$y = f(x) \quad \Rightarrow \quad \sigma_y \approx \sqrt{\left(\frac{df}{dx_1}\sigma_{x_1}\right)^2 + \left(\frac{df}{dx_2}\sigma_{x_2}\right)^2}$$

3.4 Weighted average

Have multiple measurements x_i with different errors σ_i. Want to form a weighted average with weights w_i.

$$x = \frac{\sum_{i=1}^{N} w_i x_i}{\sum_{i=1}^{N} w_i}$$
$$\sigma_x = \frac{\sum_{i=1}^{N} w_i \sigma_{x_i}^2}{\sum_{i=1}^{N} w_i}$$

Want to form an average that gives the measurements with a smaller error a larger weight, such that the error of the average is minimized. Choose $w_i = \frac{1}{\sigma_i^2}$

$$x = \frac{\sum_{i=1}^{N} x_i/\sigma_{x_i}^2}{\sum_{i=1}^{N} 1/\sigma_{x_i}^2}$$
$$\sigma_x = \frac{1}{\sum_{i=1}^{N} 1/\sigma_{x_i}^2}$$

3.5 Linear regression

Have a model, $y = mx + b$ and unknown parameters m and b. Have N measurements $y_i$with equal standard deviation σ_{y_i}at known (without error) positions x_i.

$$\hat{m} = \frac{\mathrm{Cov}(y,x)}{\mathrm{Var}(x)} = \frac{\sum_{i=1}^{N}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}$$
$$\hat{b} = \bar{y} - \hat{m}\bar{x}$$

For the errors:

$$\sigma_{\hat{m}} = \sqrt{\frac{\frac{1}{N-2}\sum_{i-1}^{N}\sigma_{y_i}^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}}$$
$$\sigma_{\hat{b}} = \sigma_{\hat{m}}\sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$$

Can obtain these errors, apart from the $\frac{1}{N-2}$ factor using the error propagation in Section 3.3.

Goodness of fit,χ^2

Assume that we know the variance of the measurement error, then

$$\chi^2_{\text{red}} = \frac{\chi^2}{\nu} = \frac{1}{N-2}\sum_{i=1}^{N}\frac{(y_i - \hat{m}x_i - \hat{b})^2}{\sigma^2_{y_i}}$$

and ν denotes the degrees of freedom$\nu = N-2$, the number of measurements minus two estimated parameters. This should be around one if the linear regression fits the data, much greater if it doesn't and much smaller values could mean that the measurements were not independent or that we were overfitting the data.

4 Parametric tests

Example t-test.

Procedure

- Formulate null hypothesis H_0:
  - The means μ_x, μ_y of two Gaussian distributions are the same.
- Compute the distribution of a test variable under the assumption that $H_0$is true.
  - Compute empirical means $\hat{\mu}_x, \hat{\mu}_y$for N_x, N_y measurements x_i, y_i.

$$\hat{\mu}_z = \frac{1}{N_z}\sum_{i=1}^{N_z} z_i, \qquad z \in \{x, y\}$$

  - Compute empirical variances $\hat{\sigma}^2_x, \hat{\sigma}^2_y$ if $\sigma^2$is unknown.

$$\hat{\sigma}^2_z = \frac{1}{N_z - 1}\sum_{i=1}^{N_z}(z_i - \hat{\mu}_z)^2, \qquad z \in \{x, y\}$$

  - Compute the weighted mean

$$\hat{S} = \frac{1}{N_x + N_y - 2}\left((N_x - 1)\hat{\sigma}^2_x + (N_y - 1)\hat{\sigma}^2_y\right)$$

and the standard error

$$S = \hat{S}\left(\frac{1}{N_x} + \frac{1}{N_y}\right)$$

– Then, if H_0 is valid,

$$t = (\hat{\mu}_x - \hat{\mu}_y)/S$$

is t-distributed with $N_x + N_y - 2$ degrees of freedom.

Dilemma I

Test questions whether a given point belongs to that distribution.

Problem This cannot be falsified. Under H_0, every value of the test statistic (t in case of the t-test) can occur.

p-value Probability that a value is greater than t:

$$p = 1 - \int_{-\infty}^{t} p(x)dx$$

If H_0 is valid, than p-value of the test statistic is uniformly distributed in $[0, 1]$.

Significance level α

Solution: Reject the null hypothesis for extreme values by defining a significance level α. Usually 0.05 or 0.01. Can reject extreme values on one (one sided) or both sides (two sided) of the distribution. Can make two kinds of errors:

- Type I error: Null hypothesis is rejected, even though it was true (false positive).
  - frequency of type I errors $< \alpha$: test is conservative.
  - frequency of type I errors $> \alpha$: test is garbage.
- Type II error: Null hypothesis is not rejected even though it was true (false negative).

Power of the test

How often $H_0$is rejected, when H_0 was false (correct rejections).

Dilemma II

Dichotomy of Kakutani