Quantitative skills for genomic data analysis (2): statistics

BBMS 3009: Genome Science (First Semester, 2021)

Dr. Yuanhua Huang / 黃淵華
School of Biomedical Sciences &
Department of Statistics and Actuarial Science

香 港 大 學
THE UNIVERSITY OF HONG KONG

Today's learning objectives (Statistics)

1. **Main idea of statistical hypothesis testing**
2. **Regression based test**
3. **Types of errors and evaluation metrics**

4. More testing methods: t-test, permutation test, etc

Resource: Chapter 6 in book *Modern Statistics for Modern Biology*

https://www.huber.embl.de/msmb/Chap-Testing.html

R script: Moodle / Lecture handouts / Dr. YH Huang → R-statistics.Rmd

1. Example of hypothesis testing

- Hypothetic example:
  - Does advertising in newspaper have impact on the sales of cars

- Decision to make
  - Yes or No

- Factors to consider when using collected data:
  - Uncertain measurement
  - Sample size is not big enough to represent the whole population

Decision making and hypothesis testing

- Common scientific questions from observed data
  - Difference between groups
  - Tendency, e.g., along with drug doses (effects)

- How to make the decision with considering
  - Uncertainty in observations, e.g., patient responses
  - Sample size is not big enough to represent the population, e.g., in clinical trial

- Statistical hypothesis testing
  - An approach for decision making under uncertainty
  - Estimate the probability to be wrong
  - Maximize expected utility (subjective value)

Main idea of hypothesis testing

A statistical hypothesis, sometimes called confirmatory data analysis, is a hypothesis that is **testable** in the light of observed data that is modeled via a set of random variables.

Main idea

- It is difficult to prove that a fact is "right".
- But it is easy to prove that an opposite fact is "wrong".
- Then you only have to find one counter example.

Main idea of hypothesis testing

- It is difficult to prove that a fact is "right", but it is easy to prove that an opposite fact is "wrong".

Research hypothesis (alternative hypothesis)

- H_1: the newspaper adverting has impact on sales

$$H_1: y = \beta_0 + \beta_1 \times \text{Newspaper}; \beta_1 \neq 0$$

Null hypothesis (default hypothesis, you don't need to prove it, just assume it)

- H_0: the newspaper adverting has no impact on sales

$$H_0: y = \beta_0 + \beta_1 \times \text{Newspaper}; \beta_1 = 0$$

Main idea of hypothesis testing

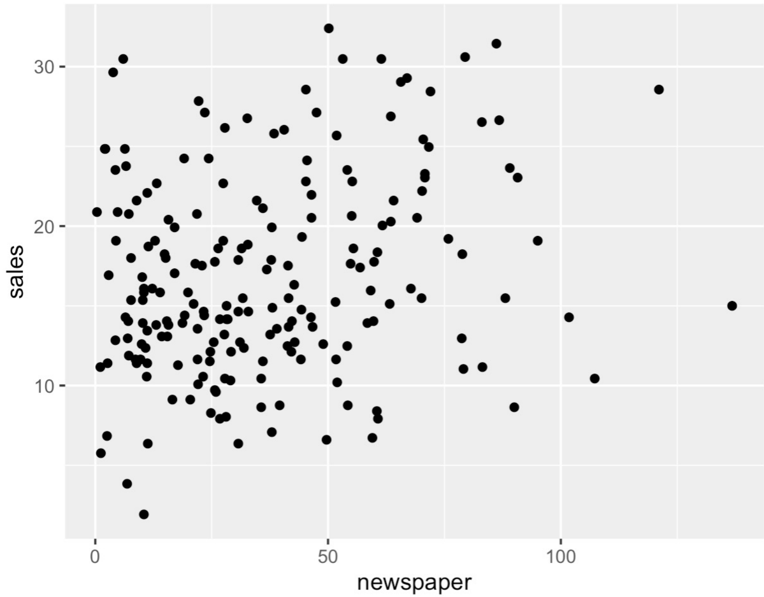
- With null and alternative hypotheses set up, we then try to show that, in light of our collected data, the null hypothesis is false.

- We do this by calculating the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct
  - If this probability is very small, it suggests that the null hypothesis is false.
  - If this probability is large, it suggests that there is not enough evidence to reject the null hypothesis.

- This probability is called the *p* value of the test

2. Regression-based testing

- Data collection
  - 200 samples with both newspaper advertising costs and sales values



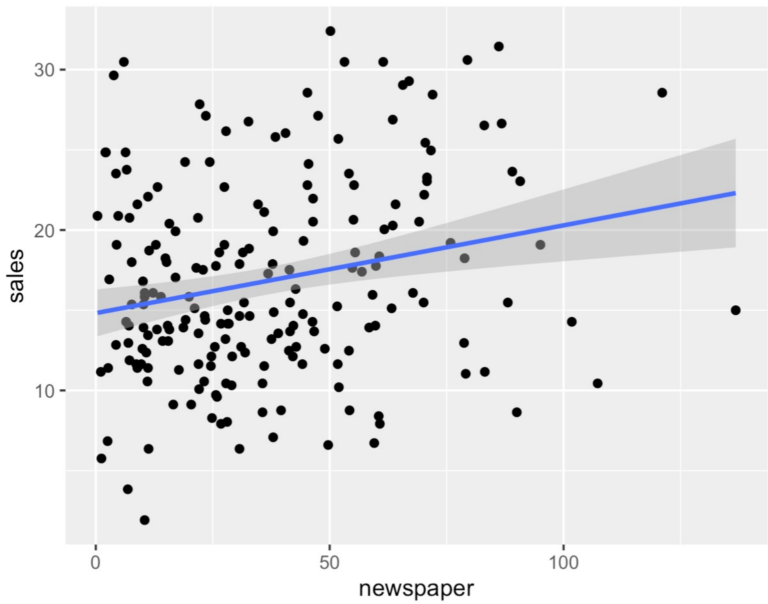

```
library(datarium)
head(marketing)

# plotting
ggplot(marketing,
aes(x=newspaper, y=sales)) +
  geom_point() +
  geom_smooth(method=lm)
```

Dataset: https://search.r-project.org/CRAN/refmans/datarium/html/marketing.html

Regression-based testing

- Fitting a regression model with maximum likelihood
  - $y = \beta_0 + \beta_1 \times \text{Newspaper};$

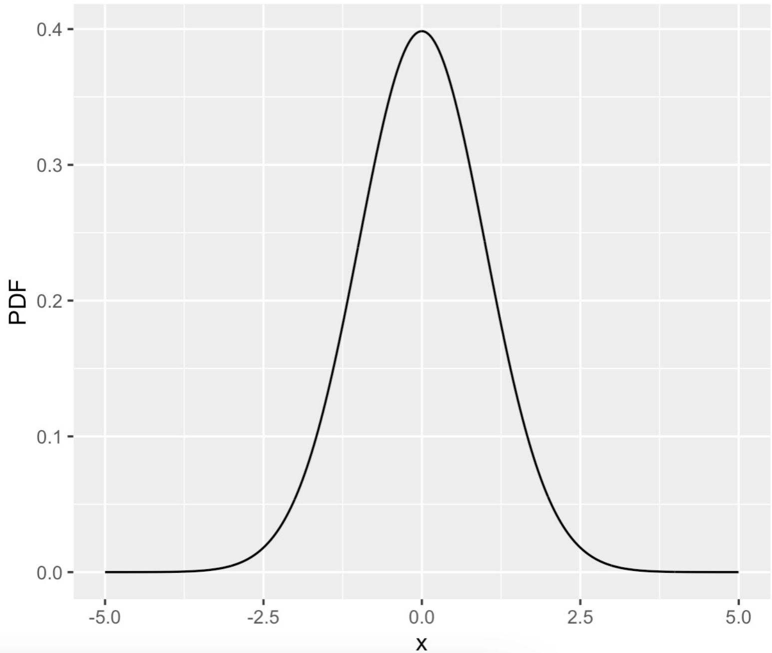


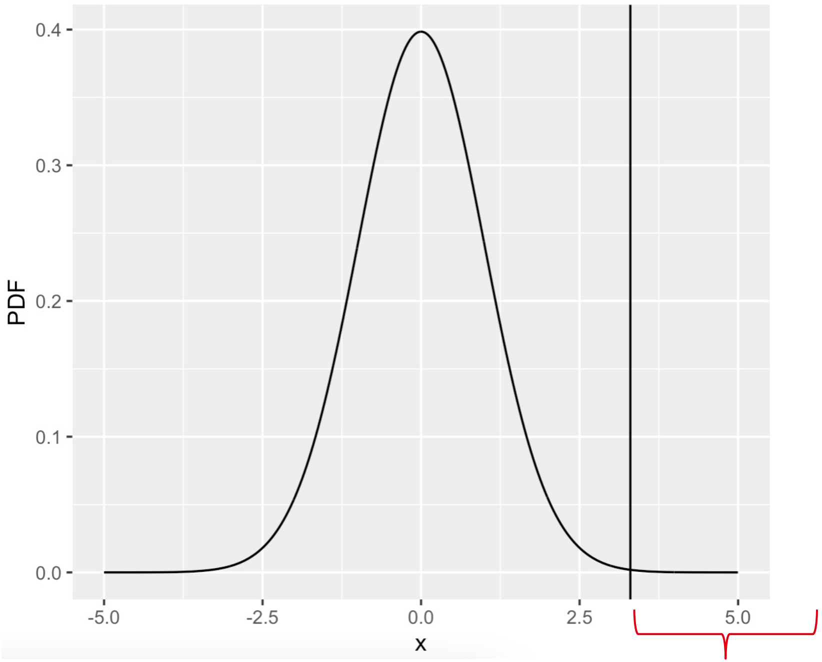
Maximum likelihood estimate: mean and standard error

Intercept: $\beta_0 = 14.82 \pm 0.746$
Newspaper: $\beta_1 = 0.0547 \pm 0.0166$

T-statistic for β_1:
t value = 0.0547 / 0.0166 = 3.3

Regression-based testing





Under the null, the distribution of t-statistic;
Degree of freedom=n_sample – n_coefficient = 198

P value = prob(x > t value) = 0.00115

P value and level of significance

- *P* value: under the null hypothesis, the probability to see the test results as extreme as observed test result
  - If *p* value is very small, it suggests that the null hypothesis is false.
  - But how low does this probability has to be before we can conclude that the null hypothesis is false?

- Convention: choose a **level of significance** before the experiment that dictates how low the *p* value should be before we reject the null hypothesis.
- It is common to choose a significance level of 0.01 or 0.05.

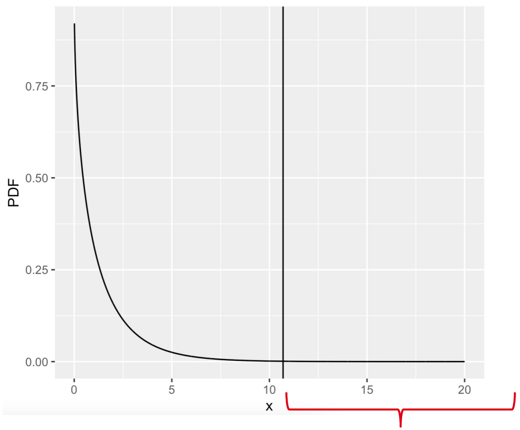
- Here $p = 0.00115$, so we reject the Null hypothesis at the significance level of 0.01.

More on regression-based testing (1)

- Likelihood ratio test
  - Null model likelihood L_0: $y = \beta_0$
  - Alternative model likelihood L_1: $y = \beta_0 + \beta_1 \times \text{Newspaper}$

- Likelihood with maximum likelihood estimate
  - Null hypothesis: $L_0 = -650.15$
  - Alternative hypothesis: $L_1 = -644.8$

- Likelihood ratio statistic
  - Observed results: $\lambda = -2(L_0 - L_1)$ = 10.7
  - Distribution under the Null: $\lambda \sim \chi^2\ (df = 1)$
  - P value: $P(x > \lambda) = 0.00107$

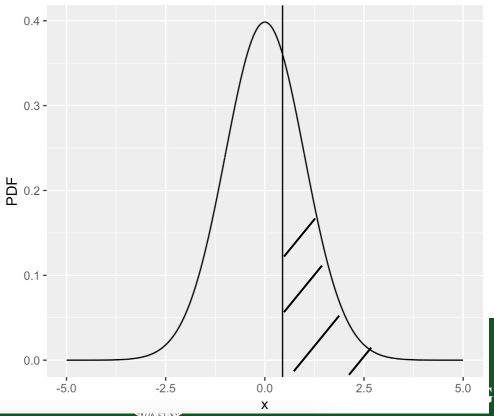


More on regression-based testing (2)

- Condition on other covariate, e.g., advertising on Facebook
  - $H_1: y = \beta_0 + \beta_1 \times \text{Newspaper} + \beta_2 \times \text{Facebook}\ ;\ \beta_1 \neq 0$
  - $H_0: y = \beta_0 + \beta_1 \times \text{Newspaper} + \beta_2 \times \text{Facebook}\ ;\ \beta_1 = 0$

	youtube <dbl>	facebook <dbl>	newspaper <dbl>	sales <dbl>
1	276.12	45.36	83.04	26.52
2	53.40	47.16	54.12	12.48
3	20.64	55.08	83.16	11.16
4	181.80	49.56	70.20	22.20
5	216.96	12.96	70.08	15.48
6	10.44	58.68	90.00	8.64

- Fitting the model with collected data
  - $\beta_0 = 11.02 \pm 0.753$
  - $\beta_1 = 0.0066 \pm 0.0149$; t value = 0.0066/0.0149 = 0.446
  - $\beta_2 = 0.199 \pm 0.022$



- P value = 0.656; fail to reject the null hypothesis at significance level of 0.01.

3. Types of errors

- Now, let's think about genome and we are testing if a gene expression is different between treatment and control for a cancer patient
- We will have a hypothesis testing for each gene, so around 10,000 tests in total.

- What errors on our decision?
  - False positive (type I error): Genes are **genuine not different**, but we thought they are (reject the null hypothesis)
  - False negative (type II error): Genes are **genuine different**, but we missed it (we didn't reject the null hypothesis)

Evaluation metrics

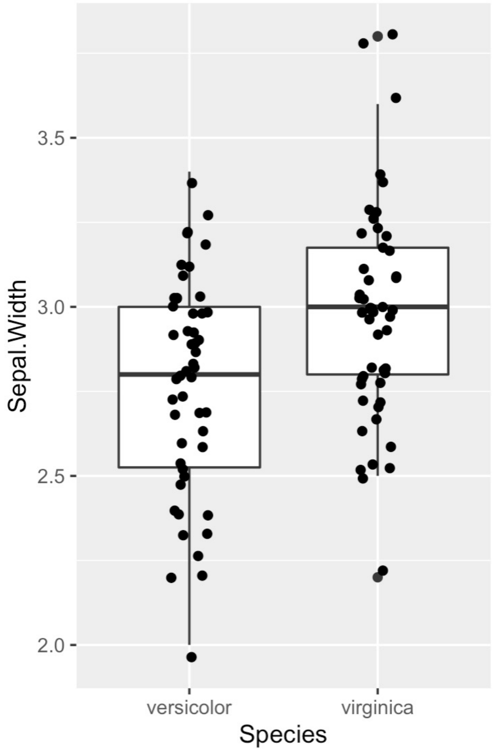
- True positive rate (Power, Sensitivity, Hit rate, Recall): $TPR = \frac{TP}{TP+FN}$
- True Negative Rate (Specificity): $TNR = \frac{TN}{TN+FP}$
- Precision (Positive Predictive Value; 1- false discovery rate):

$$Precision = \frac{TP}{TP + FP} = 1 - FDR$$

		Predicted condition	
	Total population = P + N	**Positive (PP)**	**Negative (PN)**
Actual condition	**Positive (P)**	**True positive (TP), hit**	**False negative (FN), type II error, miss, underestimation**
	Negative (N)	**False positive (FP), type I error, false alarm, overestimation**	**True negative (TN), correct rejection**

4. More hypothesis testing methods

- Is the Sepal width different between versicolor and virginica?
- Difference between groups
  - T test
  - Permutation test
  - Wilcoxon rank-sum test / Mann–Whitney U test
- These won't be examined but can be very useful for your future studies or self-learning



Resources

- Chapter 6 in book *Modern Statistics for Modern Biology*
  - https://www.huber.embl.de/msmb/Chap-Testing.html

- R script: Moodle / Lecture handouts / Dr. YH Huang → R-statistics.Rmd