Introduction to statistical tests

He Zhang
He.Zhang@UTSouthwestern.edu
Bioinformatics Core Facility
7/21/2017

Statistical tests

- Definition
  - A mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to ‘reject’ a conjecture or hypothesis about the process
- Example
  - Assess an experiment parameter
  - Compare two or more groups

Check a coin

- Observation: 6 head out of 6 coin tosses

- Question: is this coin fair?

Biology experiments

Treatment 1	Treatment 2	Treatment 3	Treatment 4
40.6	45.2	44.3	41.2
43.1	43.5	46.5	45.3
42.5	42.9	48.2	41.7
44.0	44.6	45.7	42.3
42.6	45.1		42.5

- Experiment settings
  - 19 mice, 4 treatment
- Question: do treatments affect weight?

- To answer these questions, we will use statistical hypothesis testing

- In your own researches, you often need to propose your own hypothesis

Hypothesis pair in statistical test

- Null hypothesis
  - The commonly accepted fact
  - The null hypothesis is generally assumed to be true until evidence indicates otherwise

- Alternative hypothesis
  - The opposite of the null hypothesis

How to do statistical test?

- Formulate your hypothesis
- Set α (or confidence level)
- Calculate the test statistic
- Construct Acceptance / Rejection regions
  - Or estimate p-value
- Draw a conclusion about null hypothesis

Procedure

- Step 1 - Formulate your hypothesis
  - Null hypothesis (H0)

    e.g. coin is fair ($P = 0.5$)
  - Alternative hypothesis (H1)

    e.g. coin is not fair ($P \neq 0.5$)

Procedure

- Step 2 - Set $\alpha = 0.05$

  α = Type-I error = false positive = 0.05

  Confidence level = $1 - \alpha = 0.95$

Procedure

- Step 3 – Calculate the test statistic
  - Under the null hypothesis, we can calculate a test statistic
  - e.g. Number of heads, N, follows a binomial distribution with parameter P = 0.5

    $N \sim \text{Binomial}(6, 0.5)$

Procedure

- Step 4 - Estimate the p-value
  - P-value is the probability of observing the same or more extreme events as the current observation under null hypothesis
  - e.g.

    P-value = Prob(N >= 6) = 0.5^6 = 0.015625

Procedure

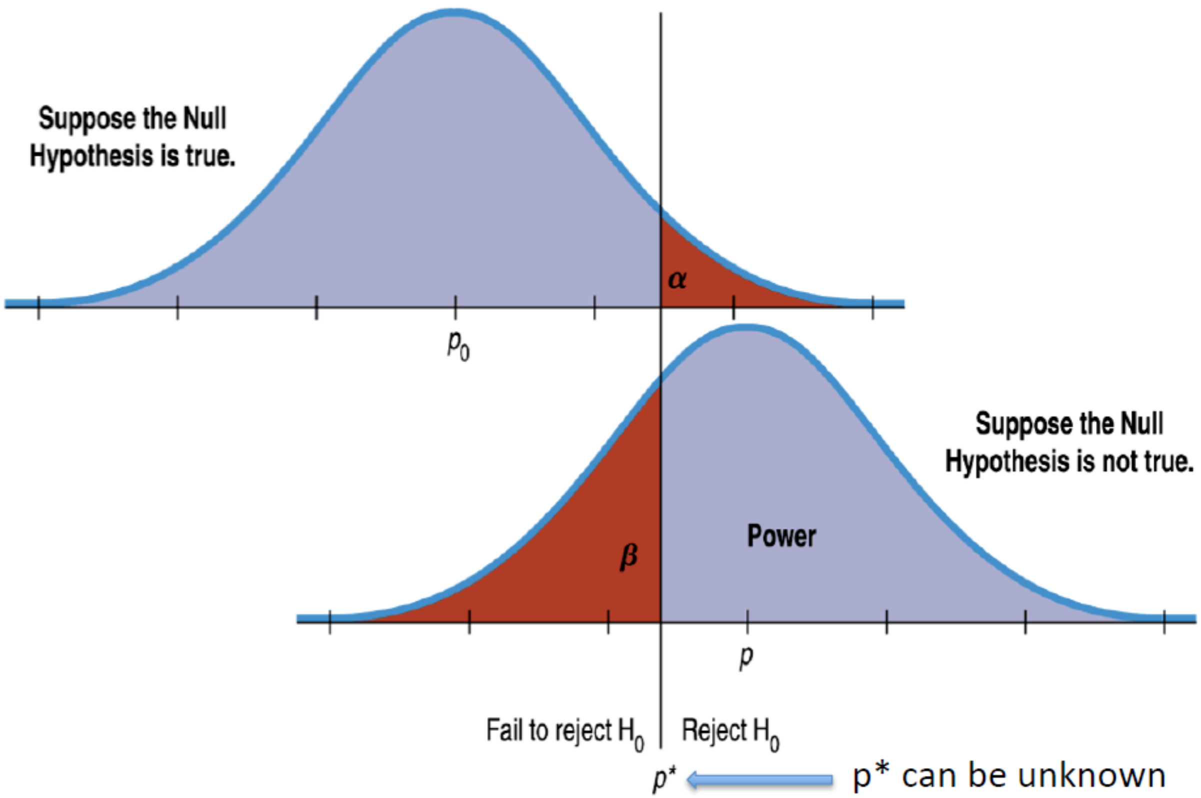
- Step 5 - Draw a conclusion
  - Compare p-value with the threshold value
    - e.g. $0.015625 < 0.05$
  - Accept/reject the null hypothesis
    - e.g. reject the null hypothesis ($P = 0.5$)
    - At the $\alpha=0.05$ level, the coin is not fair

Revisit

- Why do we need the null hypothesis
  - Null hypothesis usually represents the common view of the problem
  - Null hypothesis helps to construct the test statistics
  - Rejecting null hypothesis requires support from the data

Visualization

- Distribution of test statistics



Type I and type II errors

- Type I error
  - The null hypothesis (H0) is true, but is rejected

$$\alpha = P\{reject\ H_0 \mid H_0\ is\ true\}$$

- Type II error
  - The null hypothesis (H0) is false, but fails to be rejected

$$\beta = P\{accept\ H_0 \mid H_0\ is\ false\}$$

- Power
  - The null hypothesis (H0) is false, and is rejected

$$\pi = 1 - \beta = P\{reject\ H_0 \mid H_0\ is\ false\}$$

Interpretation

- P-value
  - NOT error rate
  - P-value is significant, any of these can happen
    - Null hypothesis is true, but observe the current data
    - Alternative hypothesis is true
- Example
  - Assuming the null hypothesis holds, the probability of observing the same or more extreme effect as the one is the data.

One-sided or two-sided test?

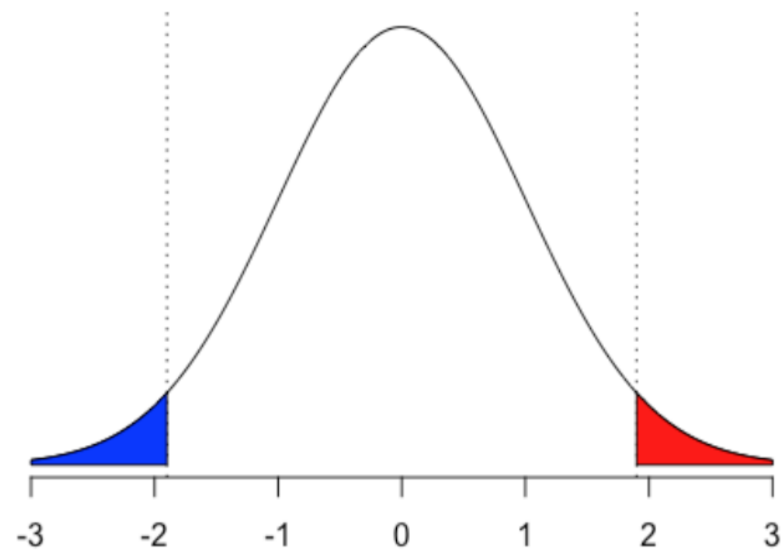
- One-sided test

  H0: $\mu = 0$ vs H1: $\mu > 0$

  H0: $\mu = 0$ vs H1: $\mu < 0$

- Two-sided test

  H0: $\mu = 0$ vs H1: $\mu \neq 0$



One-sided p-value = 0.0287
Two-sided p-value = 0.0574

One-sided or two-sided test?

- When is a one-sided test appropriate
  - You have a hypothesis about the direction of an effect
  - Missing an effect in the untested direction are negligible

- Example
  - you have developed a new drug. It is cheaper than the existing drug and, you believe, no less effective. In testing this drug, you are only interested in testing if it less effective than the existing drug.

One-sided or two-sided test?

- When is a one-sided test NOT appropriate
  - Choosing a one-sided test for the sole purpose of attaining significance is not appropriate.
  - Choosing a one-sided test after running a two-tailed test that failed to reject the null hypothesis is not appropriate, no matter how "close" to significant the two-tailed test was.

T-test

- One-sample test
  - whether the mean of a population has a value specified in a null hypothesis
  - e.g. whether the mean weight of a group of mice is equal to 45g

- Two-sample test
  - whether the means of two populations are equal
  - e.g. whether the mean weight of two groups of mice are equal.

- Two-sample paired test
  - the difference between two responses measured on the same statistical unit has a mean value of zero
  - e.g. for the same group of mice, whether the mean weight changed after a treatment.

- Test of beta in regression
  - whether the slope of a regression line differs significantly from 0

Assumptions of t-test

- Continuous variable
- Independence
- Normality
- Same population variances (homogeneous)

T-test

- Is there significant difference in LDL levels between men and women
  - H_0: Difference between LDL levels in men and women are 0
  - H_1: Difference between LDL levels in men and women are not 0

t.test(x.male$LDL, x.female$LDL)

```
> t.test(x.male$LDL, x.female$LDL)

        Welch Two Sample t-test

data:  x.male$LDL and x.female$LDL
t = -6.2186, df = 871.91, p-value = 7.772e-10
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.4579649 -0.2382342
sample estimates:
mean of x mean of y
 2.757002  3.105102
```

T statistic

H_1

Confidence level (1-α)

Wilcoxon rank-sum test

- Nonparametric test
  - Not based on parameterized families of probability distributions
  - it does not require the assumption of normal distributions

- Alternative to the two-sample t-test

- Other names
  - Mann–Whitney U test
  - Mann–Whitney–Wilcoxon (MWW)
  - Wilcoxon–Mann–Whitney test

Assumptions

- Independence
- The responses are ordinal
  - For any two observations, you can tell which is the greater one.

Wilcoxon rank-sum test

- Is there significant difference in LDL levels between men and women ?
  - H_0: Difference between LDL levels in men and women are 0
  - H_1: Difference between LDL levels in men and women are not 0

wilcox.test(x.male$LDL, x.female$LDL)

```
> wilcox.test(x.male$LDL, x.female$LDL)

        Wilcoxon rank sum test with continuity correction

data:  x.male$LDL and x.female$LDL
W = 91848, p-value = 4.272e-10
alternative hypothesis: true location shift is not equal to 0
```

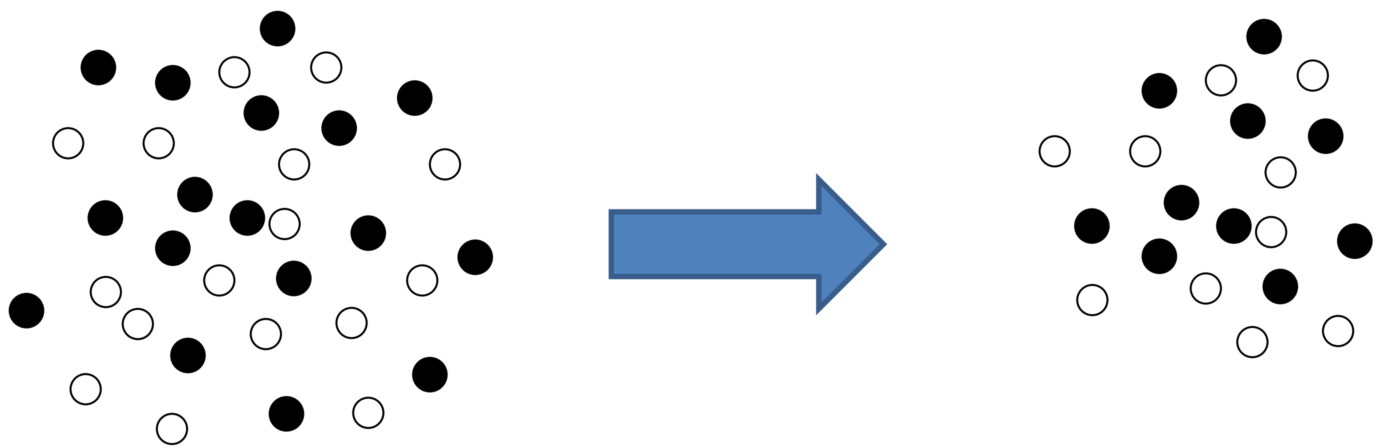
Hypergeometric distribution

- Discrete probability distribution
- Describe the probability of k successes in n draws, without replacement, from a finite population of size N that contains exactly K successes, wherein each draw is either a success or a failure.

Hypergeometric distribution

Population

Draw



Total: N = 100 balls
White: K = 30 balls

Draw: n = 20 balls
White: k = ? (0, 1, 2, …, 20)

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Hypergeometric test

- Hypergeometric test can be used in enrichment test

- Experiment (e.g. RNA-seq to identify differentially expressed genes)
  - 20,000 (N) genes were assayed
  - 100 (K) genes were from Pathway A
  - 150 (n) differentially expressed genes were identified
  - 20 (k) differentially expressed genes were from Pathway A

- Question
  - Are the genes from Pathway A significantly enriched in the differentially expressed genes?

Hypergeometric test

- Hypothesis
  – H0: Genes from Pathway A were not enriched
  – H1: Genes from Pathway A were enriched

- Calculation of P-value
  – P-value = P(X=20) + P(X=21) + … + P(X= 150)

- R code
  – phyper(k - 1, K, N - K, n, lower.tail = F)
  – phyper(20 - 1, 100, 20000 - 100, 150, lower.tail = F)

```
> phyper(20 - 1, 100, 20000 - 100, 150, lower.tail = F)
[1] 2.770949e-23
```

Fisher's exact test

- The test is useful for categorical data that result from classifying objects in two different ways.
- Examine the association (contingency) between the two kinds of classification.
- Example
  - Is there association between gender and cardiovascular disease (CAD)?

CAD	Female	Male
No	493	290
Yes	43	174

Fisher's exact test

- Hypothesis
  - H0: There is no association between gender and CAD
  - H1: There is association between gender and CAD

- R code
  - fisher.test(x.cad.sex)

Fisher's exact test

```
> fisher.test(x.cad.sex)

        Fisher's Exact Test for Count Data

data:  x.cad.sex
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  4.730158 10.135358
sample estimates:
odds ratio
  6.865069
```

- Odds ratio (OR) in fisher.test

	1	2
A	A1	A2
B	B1	B2

$$\text{odds ratio} = \frac{\frac{B2}{A2}}{\frac{B1}{A1}} = \frac{\frac{B2}{B1}}{\frac{A2}{A1}} = \frac{B2 * A1}{B1 * A2}$$

Pearson Correlation

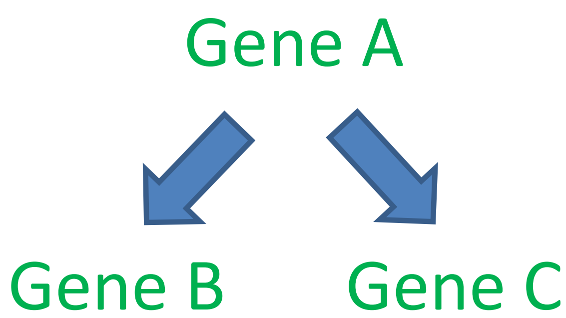
- Pearson correlation coefficient (r) is a measure of the linear correlation between two variables X and Y

- $-1 \leq r \leq 1$
  - r = 0 , no linear association; does not imply independence
  - r = 1, fit perfectly by an increasing line
  - r = -1, fit perfectly by an decreasing line

Pearson Correlation

- Correlation can not answer questions of causality directly

Correlation ≠ Causality



The expression levels of B and C may be correlated.
B or C is not directly regulated by each other.

Summary

- Statistical hypothesis (H0 and H1)
- Procedure of hypothesis test
- Interpretation of hypothesis testing result
- Commonly used hypothesis tests