Biostatistics and R

Wanwei Zhang, PhD

Department of Microbiology & Immunology

HHSC 910

wz2370@cumc.columbia.edu

In this course, you'll learn …

- the basic theory of statistics
  - Sample mean, standard deviation, distribution, statistical test
- the basic skills of statistical analysis in R language.

Random variable

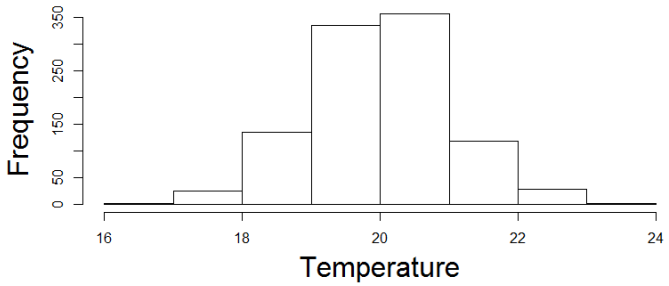
- Discrete variable
  - Weather: sunny, rainy, cloudy, …
- Continuous variable
  - Temperature: 20, 20.9, 20.99, …
- Denotation
  - Upper-case letter (X, Y, Z, …) denotes a random variable
  - Lower-case letter (x, y, z, …) denotes a specific value of a variable
- Event
  - A random variable takes a specific value or set of values
    - $X = x$
    - $X > x$
- Sample
  - A series of observations of a random variable

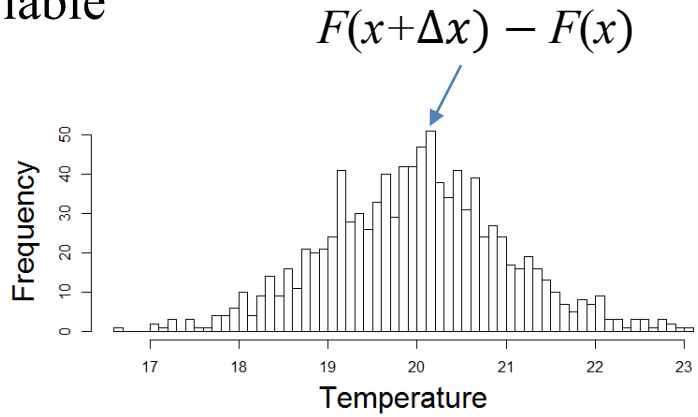
Probability and distribution of discrete variable

- Probability measures the likelihood of an event to occur
  - P(weather is sunny)
  - $P(X = x)$
- Probability density function (p.d.f.) of a discrete variable is the collected probabilities of all possible values of that variable
  - $f(x)$, $p(x)$
- Each kind of p.d.f. defines a type of distribution

Continuous variable?

- P(X = x) = 0 for any x!
- Discretization of continuous variable





- Cumulative distribution function (c.d.f) F(x)

$$F(x) = \mathrm{P}(\mathrm{X} \geq x)$$

- Probability density function (p.d.f.) of a continuous variable can be defined as the derivative of c.d.f

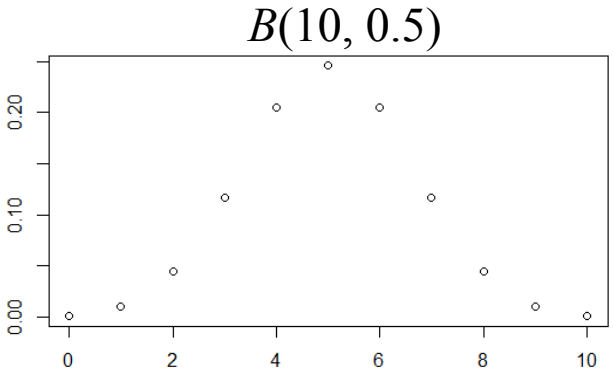
$$f(x) = F'(x) = \lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x}$$

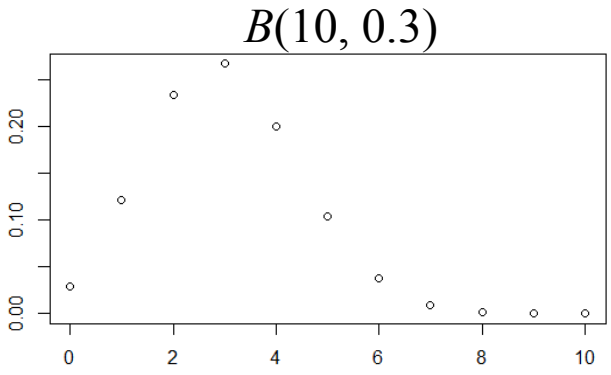
Common distributions

- Discrete
  - binomial distribution
  - hypergeometric distribution
  - Poisson distribution
  - …
- Continuous
  - Normal/Gaussian distribution
  - Student's t-distribution
  - …

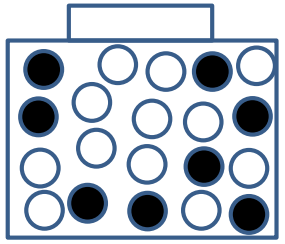
Binomial distribution

- Example 1: flip a coin 10 times, then the number of heads you get follows binomial distribution $B(10, 0.5)$





- Example 2: randomly draw 6 balls, one by one with replacement, then number of white balls follows binomial distribution $B(6, \frac{12}{20})$

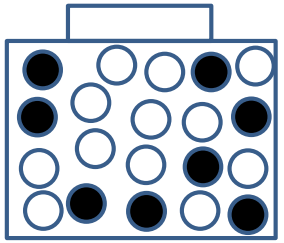


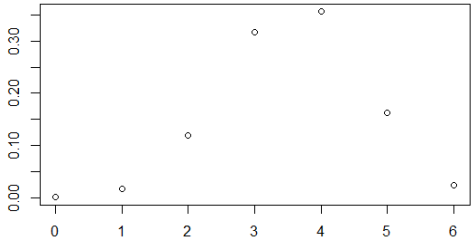
white: 12
black: 8

$$p(x; n, p) = \binom{n}{x} p^x (1-p)^{n-x}$$

Hypergeometric distribution

- Example 1: Randomly draw 6 balls from jar **without replacement**, then number of white balls follows the hypergeometric distribution $H(6, 12, 20)$

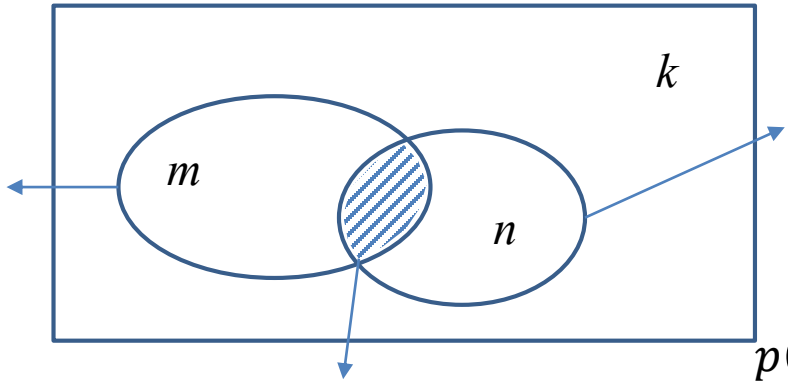




$H(6, 12, 20)$

- Example 2:

Total # of human genes



# of differentially expressed genes

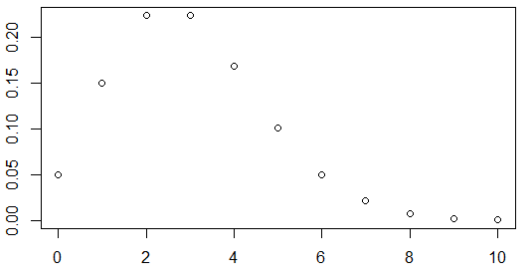
# of genes in a pathway

# of differentially expressed genes in the pathway follows $H(n, m, k)$

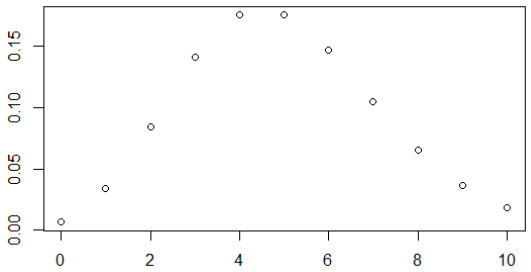
$$p(x; n, m, k) = \frac{\binom{n}{x}\binom{k-n}{m-x}}{\binom{k}{m}}$$

Poisson distribution

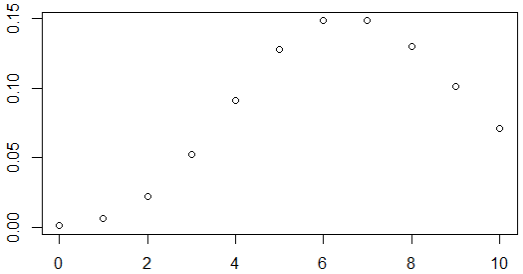
- Example 1: the amount of mails a post office receive in a given time interval follows Poisson distribution Pois(λ).



Pois(3)

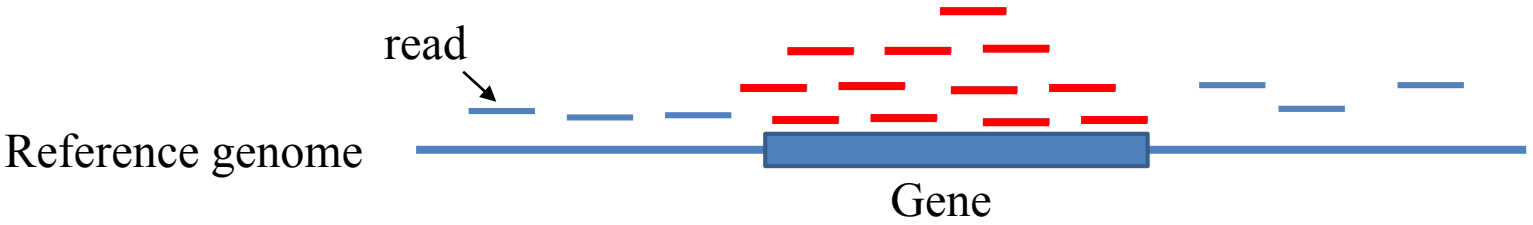


Pois(5)



Pois(7)

- Example 2:



Number of reads aligned to a gene region follows Poisson distribution

$$p(x;\lambda) = e^{-\lambda}\frac{\lambda^x}{x!}$$

Expected value and standard deviation

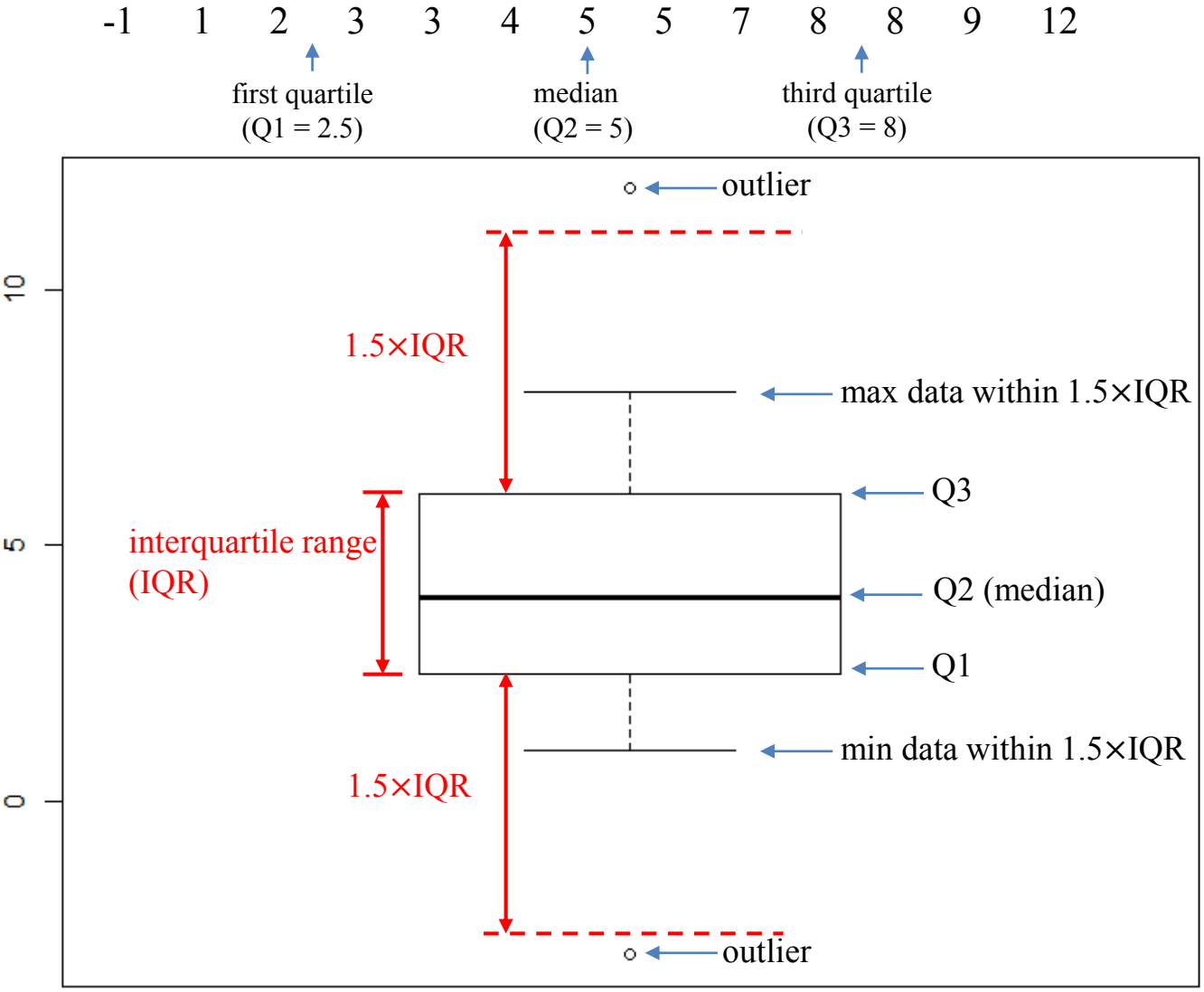
- Expected value (Population mean, μ), E
  - Discrete variable $E(X) = \sum_{i=1}^{n} x_i f(x_i)$
  - Continuous variable $E(X) = \int x f(x) dx$
- Standard deviation (σ)
  - $\sigma = \sqrt{E[(X - \mu)^2]}$

Sample mean and median

- Sample mean ($\bar{x}$)
  - $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
- Median ("middle" value)
  - {1, 2, 3, 3, 4, 5, 5, 7, 8}
  - {1, 2, 3, $\underbrace{3, 5}_{\text{average}}$, 5, 7, 8}
- Mean vs median
  - {1, 2, 3, 3, 4, 5, 5, 7, 8}
    - Mean ≈ 4.2; Median ≈ 4
  - {1, 2, 3, 3, 4, 5, 5, 7, 80}
    - Mean ≈ 12.2; Median ≈ 4
- Median is less sensitive to outliers

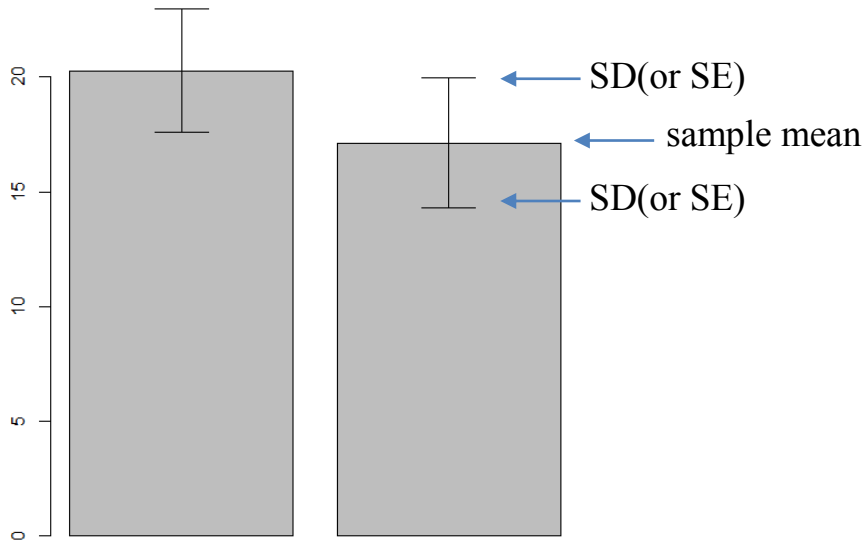
Quartiles and Boxplot

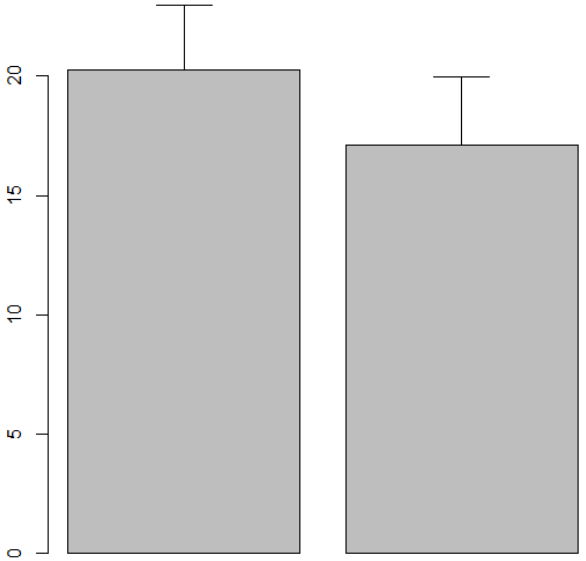
Quartiles are values that split sorted data into 4 equal parts.



Sample standard deviation and standard error of the mean

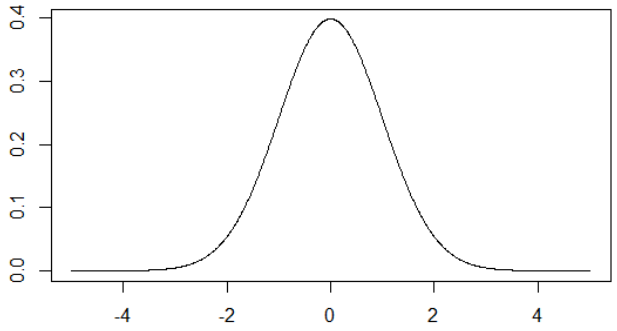
- Sample standard deviation (SD, S)
  - $S = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$ S is an unbiased estimator of σ, i.e. $E[S] = \sigma$
- Standard error (SE) of the mean
  - SE = $S/\sqrt{N}$
- Bar chart with error bar





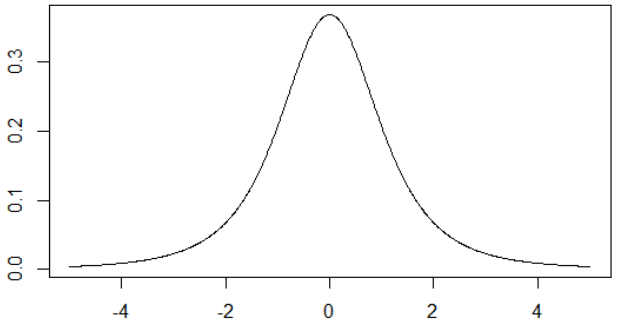
Normal distribution and student's t distribution

Normal distribution $N(\mu, \sigma)$



$N(0,1)$
standard normal distribution

student's t distribution $T(\nu)$



$T(3)$

sample mean

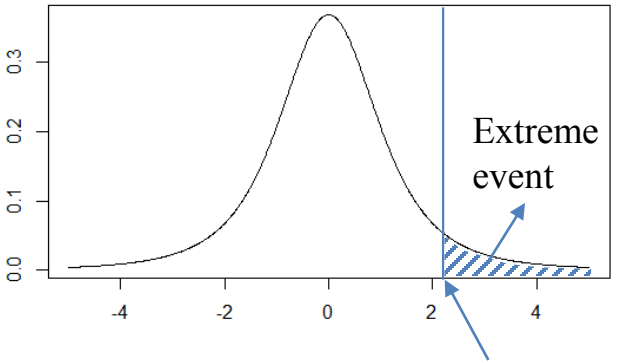
sample size

Theorem If $X \sim N(\mu, \sigma)$, then $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim T(n-1)$

sample standard deviation

Statistical inference (testing hypothesis)

- Null hypothesis (H_0)
  - Average GPA among female students equals the one among male students
  - TP53 expression in tumor equals the one in non-tumor tissue
- Alternative hypothersis (H_1)
  - Average GPA among female students is greater than the one among male students
  - TP53 expression in tumor is different from the one in non-tumor tissue
- Test statistic
  - A quantity calculated from sample and follows a definite distribution under null hyposthesis
- Extreme event
  - Event that includes test statistic and is supposed to **unlikely** happen under null hypothesis
- P-value
  - the probability that extreme event would occur under null hypothesis



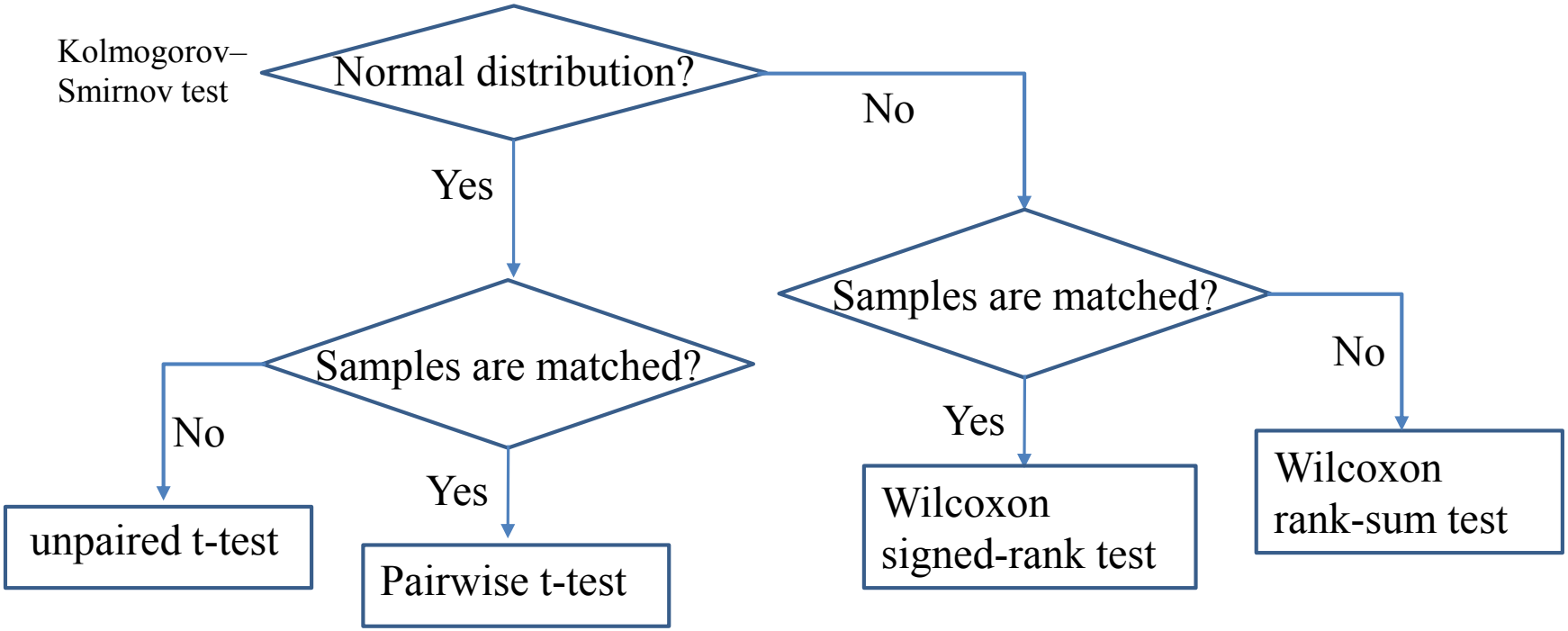
Commonly used test methods

- Student's t-test
  - Samples follow normal distribution (use Kolmogorov–Smirnov test to assess the normality)
- Wilcoxon rank-sum test (Mann–Whitney U test)
  - Samples are independent
- Wilcoxon signed-rank test
  - Samples are matched or dependent
- ANalysis Of VAriance (ANOVA)
  - Test whether or not the means of several groups are equal
  - Variance of each groups should be the same (use Bartlett's test to assess homoscedasticity)
- Fisher's exact test
  - Gene set enrichment analysis

Choosing a proper test

- What does your data's distribution look like? (Gaussian or not)
- What's the hypothesis you are testing?
- What's the pre-assumptions of the test method you choose?

Scheme for 2-sample comparison:



Multiple tests correction

- P-value tells us the probability we make a wrong decision.
- In a single test, p-value<0.05 guarantees the small probability of making a wrong decision
- Multiple tests make hundreds and thousands of decisions, which makes the chance of wrong getting much higher. (False discovery rate, FDR)
- It's necessary to do multiple tests correction to reduce the chance of making wrong decisions.
- There are many correction methods,
  - Bonferroni
  - Benjamini & Hochberg
  - …

R Practice: preparing

- Create a new directory *stat17* in your computer
- Download data from ftp://156.111.46.66/stat17
  - username: stat17
  - password: 123456
- Open Rstudio
- Open “scripts.R” in Rstudio

R Practice 1: file manipulation

- Loading data from file
  - read.table(file, header = FALSE, sep = "\t", row.names, col.names, …)
  - read.csv(file, header = TRUE, …)
- Writing data to file
  - write.table(x, file = "", sep = "\t", row.names = TRUE, col.names = TRUE)
  - write.csv(x, file = "", row.names = TRUE, col.names = TRUE)
- Examples
  - Load temperature.txt into variable “t”
  - Load lungCancerData.txt into variable “lung”

R Practice 2

- mean, median
- standard deviation, standard error of mean
- boxplot
- histogram
- barplot with error bar

R Practice 3

- Two sample t-test
- Multiple test correction
- Volcano plot
- Heatmap