Outline

Contents

1 Probability distributions

Random variables and probability

- A brief overview of random variables.
- $X(\omega) : \Omega \to \mathbb{R}$.
- Binary (Bernoulli), a special case of discrete r.v.
- Discrete (binomial, Poisson, negative binomial, ...)
- Continuous (normal, chi-squared, logistic, ...)
- Notion of independence/dependence; $i.i.d.$; what constitutes a **sample**?

Built-in probability functions

```
x1 <- rgeom(10, 1/3)        #geometric r.v., discrete.
x2 <- rbinom(10, 1, 1/2)    #standard Bernoulli (fair coin)
x3 <- rnorm(10)             #standard normal (mean=0, sd=1)
x4 <- rnorm(100, 5, 2)     #mean=5, sd=2.
mean(x4); sd(x4)            #convergence

## [1] 5.081998
## [1] 2.014264
```

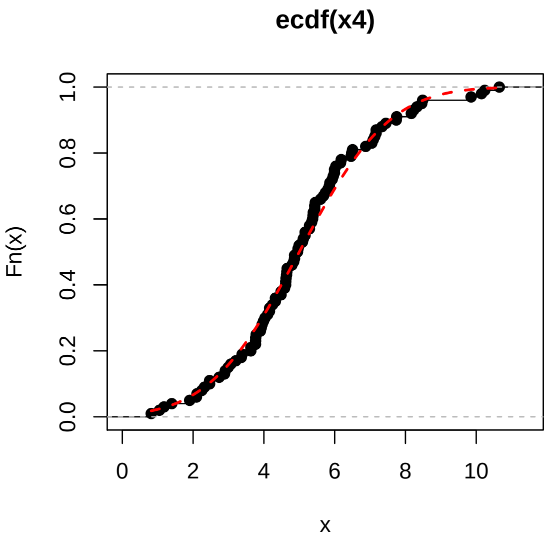
Built-in functions (II)

- Probability density function, *a.k.a. p.d.f.*. qDiscrete/continuous. In R: `dbinom()`, `dnorm()`, `dchisq()`, etc.
- Probability distribution function, *a.k.a. c.d.f.*. In R: `p<name>`.
- Quantile function. In R: `q<name>`.

Useful functions

- `summary(), fivenum(), stem()`.
- `hist(), density(), rug(), ecdf()`.
- QQ-plots: follow the book example.
- Formal normality tests. Useful in residual analysis (goodness of fit).

```
grid <- seq(min(x4), max(x4), 0.01)
plot(ecdf(x4))
lines(grid, pnorm(grid, 5, 2), lty=2, col="red", lwd=2)
```



2 Group comparisons

Quick review of hypothesis testing

- A typical scenario in data analysis is to compare groups.
- Even if there is no "true difference", sample means calculated from different groups will be different due to randomness.
- Hypothesis testing: H_0: no group effects versus H_1: group effects are not zero. p-value in a nutshell: How likely we'll end up with the observed group difference under H_0?
- Hypothesis testing can be generalized to all correlation models, regression models, etc.

One and two sample tests for continuous data

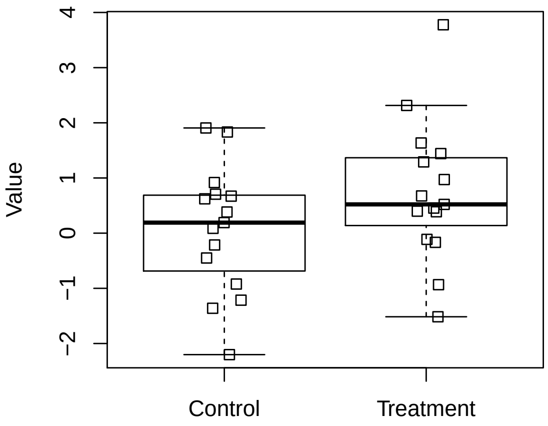
- `t.test()` and variants (one-group v. two group; one-sided v. two sided; paired v. unpaired).
- `var.test()`. The two-fold rule-of-thumb and the Welch correction.
- Parametric v. nonparametric test.
- `wilcox.test()` (paired, unpaired).
- A few words about `shapiro.test()`. "Passing" this test ($p > 0.05$) for very small sample size does not mean much. Don't do 3 mice versus 3 mice experiments!!
- Omnibus test: Kolmogorov-Smirnov test `ks.test()`, Cramer-von Mises test (`cvm.test()`), etc.

Two-sample t-test and Wilcoxon ranksum test

```
X <- rnorm(15); Y <- rnorm(15) + .8*X
t.test(X, Y)     # Default: two-sided, nonpaired, with correction.
XY <- c(X, Y)
Grp <- c(rep("Control", length(X)), rep("Treatment", length(Y)))
t.test(XY~Grp)  #the equivalent formula interface.
# one-sided test
t.test(XY~Grp, alternative="less") #less means X<Y
# Wilcoxon ranksum test is the nonparametric counterpart
## of two sample t-test
wilcox.test(XY~Grp, alternative="less")
```

Two-sample tests (II)

```
## Boxplot of the data. It is good to plot the actual data
## (jittered a little bit) on the boxplot.
boxplot(XY~Grp, outpch = NA, xlab="", ylab="Value")
stripchart(XY ~ Grp, vertical=TRUE, method="jitter", add=TRUE)
```

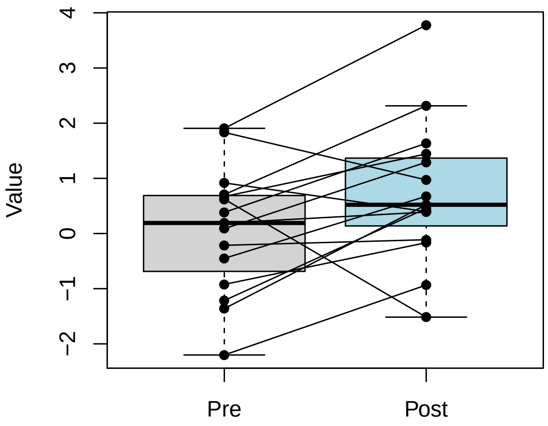


Paired tests

```r
## paired, one-sided test (Y greater than X
Grp2 <- c(rep("Pre", length(X)), rep("Post", length(Y)))
t.test(X, Y, paired=TRUE)
# Wilcoxon signed rank test
wilcox.test(X, Y, paired=TRUE)
```

Paired test (II)

```r
boxplot(X, Y, outpch = NA, xlab="", names=c("Pre", "Post"),
        ylab="Value", col=c("lightgrey", "lightblue"))
stripchart(list(X, Y), vertical=TRUE, pch=16, add=TRUE)
## Add line segments to link pre/post together
s <- seq(length(X))
segments(rep(1,length(X))[s],X[s],rep(2,length(X))[s],Y[s])
```



3 Linear models and ANOVA

Linear Regression

- Linear regression is the most popular way to model *linear* relationship between covariates ($\mathbf{X}$) and continuous responses ($\mathbf{Y}$).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

  Here $\boldsymbol{\varepsilon}$ is the error term and is usually (but not always) modeled as multivariate normal random vector.

- Statistical inference of LM includes estimating $\boldsymbol{\beta}$, provide an overall p-value for goodness-of-fit, and individual p-values for each β_k.

- Advanced topic not covered here: Model selection based on AIC/BIC or LASSO/elastic net.

Analysis of Variance for linear regression

- The null and alternative hypothesis of nested linear models `mod1` and `mod0`.
- H_0: Additional covariates in `mod1` do not have significant effect.
- H_1: Some covariates in `mod1` have significant effect.
- The F-test
  $$F = c\frac{\mathrm{RSS}_0 - \mathrm{RSS}_1}{\mathrm{RSS}_1}. \tag{2}$$
  Here c is a constant that depends on the degrees of freedom of both models.

ANOVA for multi-group comparisons

- The same variance decomposition principle can be used to analyze group-effect in multi-group comparisons.
- One-way ANOVA F-test. Nonparametric counterpart: Kruskal-Wallis test.
- Repeated measures (paired) ANOVA. Nonparametric counterpart:

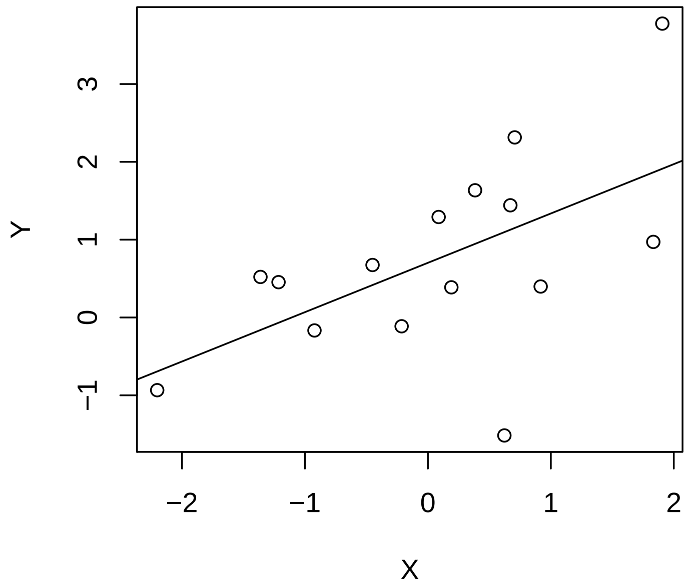
LM example

```
mod1 <- lm(Y~X)
mod0 <- lm(Y~1)    #the null model

summary(mod1)
anova(mod1, mod0) #p-value is the same as F-pvalue in mod1
```

LM plot

```
plot(Y~X)
abline(mod1)
```



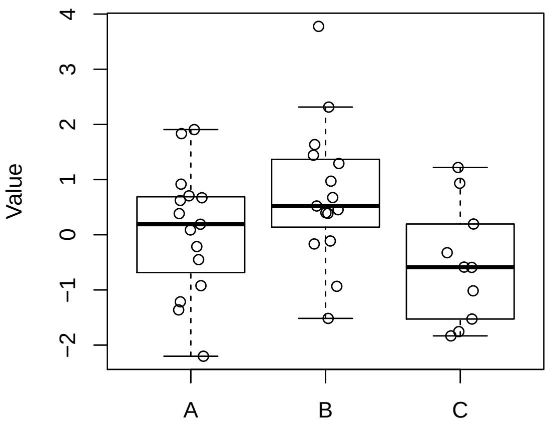
One-way ANOVA

```
Z <- rnorm(10)
XYZ <- c(X, Y, Z)
Grp3 <- factor(c(rep("A", length(X)), rep("B", length(Y)),
                 rep("C", length(Z))))

## one-way ANOVA F-test
anova(lm(XYZ ~ Grp3))
## Function aov() is a shortcut
summary(aov(XYZ ~ Grp3))

# Kruskal-Wallis test (nonparametric)
kruskal.test(XYZ, Grp3)
```

```
boxplot(XYZ~Grp3, outpch = NA, xlab="", ylab="Value")
stripchart(XYZ ~ Grp3, vertical=TRUE, method="jitter",
           add=TRUE, pch=1)
```



Post-hoc analysis

- More than often, we want to know which pairwise group comparison is significant.
- Proper way: 1. Test for overall significant. 2. Apply a suitable *post hoc* analysis which controls the overall type I error.
- Methods: Tukey's *post-hoc* analysis procedure for parametric test;
- Common mistakes: 1. Pairwise t-test without adjustment.

Post-hoc analysis example

```
## Tukey's procedure. Good for parametric test.
TukeyHSD(aov(XYZ ~ Grp3))

## Dunn's test. Good for nonparametric test.
# install.packages("dunn.test")
library("dunn.test")
## Here method="hs" means Holm-Sidak adjustment
dunn.test(XYZ, Grp3, method="hs")
```

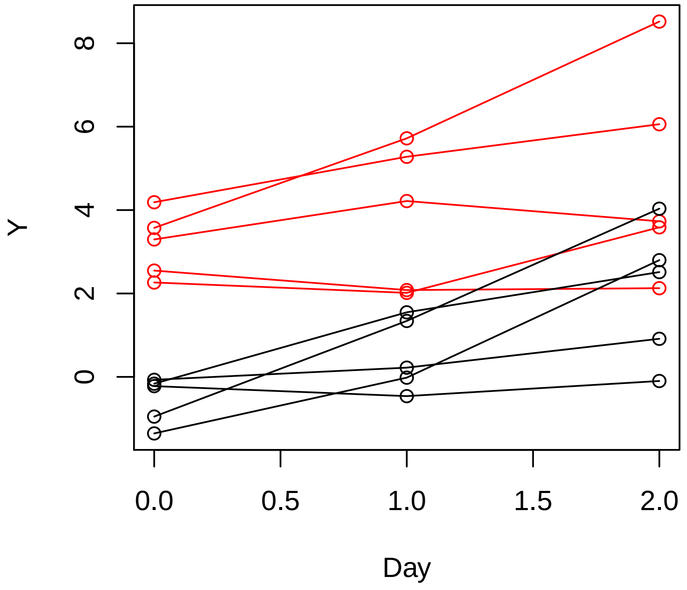
Repeated measures ANOVA

- Imaging that you are observing data collected from 10 subjects (5 girls and 5 boys) at three time points: Day 0, 1, and 2.
- You want to test whether there is a significant Day effect or a Gender effect.
- Ordinary regression or one-way ANOVA is not appropriate due to correlation between errors.
- Solution: Repeated measures ANOVA and its nonparametric counterpart, Friedman's test.

Repeated measures ANOVA example (I)

```r
## generate some longitudinal data
Gender <- rep(c("Female", "Male"), each=5)
Day0 <- rnorm(10) + ifelse(Gender=="Female", 3, 0)
Day1 <- Day0 + ifelse(Gender=="Female", 0, 1) + rnorm(10)
Day2 <- Day1 + ifelse(Gender=="Female", 0, 1) + rnorm(10)
## Subject names
SN <- paste("sub", rep(1:10, 3), sep="")
## combine them together
mydata <- data.frame(Y=c(Day0, Day1, Day2),
                     Day=rep(0:2, each=10),
                     Gender=rep(Gender, 3),
                     Subject=SN)
```

```r
plot(Y~Day, data=mydata, col=ifelse(Gender=="Female", "red", "black"))
for (i in 1:10){
    lines(Y~Day, data=mydata[mydata[, "Subject"]==paste("sub", i, sep=""),],
          col=ifelse(Gender=="Female", "red", "black"))
}
```



Repeated measures ANOVA example (II)

```r
mod2 <- aov(Y ~ Day + Error(Subject), data=mydata)
summary(mod2)
## Two-way ANOVA with
mod3 <- aov(Y ~ Day+Gender + Error(Subject), data=mydata)
summary(mod3)
```

```
## Nonparametric version in simple case
friedman.test(Y ~ Day | Subject, data=mydata)
```

Advanced linear regression and ANOVA techniques

- A full-fledged linear mixed effect model can have many fixed and random factors, with the randomness encoded in both the intercept and slope terms. Package: `lme4`, function `lmer()`.
- Robust regression. `MASS`, `rlm()`.
- Ordinal regression. `polr()`, `MASS`.

4 Logistic Regression

- What if the response variable is *binary*?
- Answer: Logistic (or probit) regression.

$$\operatorname{logit} p := \log \frac{p}{1-p} = \mathbf{X}\boldsymbol{\beta}. \tag{3}$$

- The above model is a special case of *generalized linear model*, which also includes probit regression, Poisson regression, etc.
- Function: `glm()`.

Logistic regression

```
Smoke <-  c(rep(0, 10), rep(1, 10), rep(2, 10), rep(3, 10))
Cancer <- c(rep(0, 10), rbinom(10, 1, .3), rbinom(10, 1, .5), rbinom(10, 1, .8))
mod6 <- glm(Cancer ~ Smoke, family=binomial(link=logit))
summary(mod6)
```

Other advanced regression models

- Predicting expected rates of counting data in Poisson regression. Again `glm()`, with link function set to Poisson.
- GLMs can also have random effect. Use `glmer()` from the `lme4` package.
- Nonparametric regression, additive model, etc.

5 Introduction to categorical data analysis

$p \times q$ Contingency table

- The association between smoking (as a binary variable) and lung cancer.
- Once summarized, it is a 2×2 table.
- Suitable statistical test: χ^2-test (Chi-square test, an approximate parametric test) and Fisher's exact test (non-parametric).

Example of 2×2 Contingency table

```
Smoke.binary <- ifelse(Smoke==0, 0, 1)
ctab1 <- table(Smoke.binary, Cancer)
ctab1
ctab2 <- table(Smoke, Cancer)         #4x2 table
chisq.test(ctab1)

## Warning in chisq.test(ctab1):  Chi-squared approximation may be incorrect

fisher.test(ctab1)
chisq.test(ctab2)

## Warning in chisq.test(ctab2):  Chi-squared approximation may be incorrect

fisher.test(ctab2)
```

Generalized Cochran-Mantel-Haenszel Tests

- Cochran-Mantel-Haenszel test (function `mantelhaen.test()`) can test $p \times q$ table observed at several different timen points.
- Package `vcdExtra` has a function `CMHtest()` that can test the association between two *ordinal* factors, possibly observed at several time points.

6 Other topics

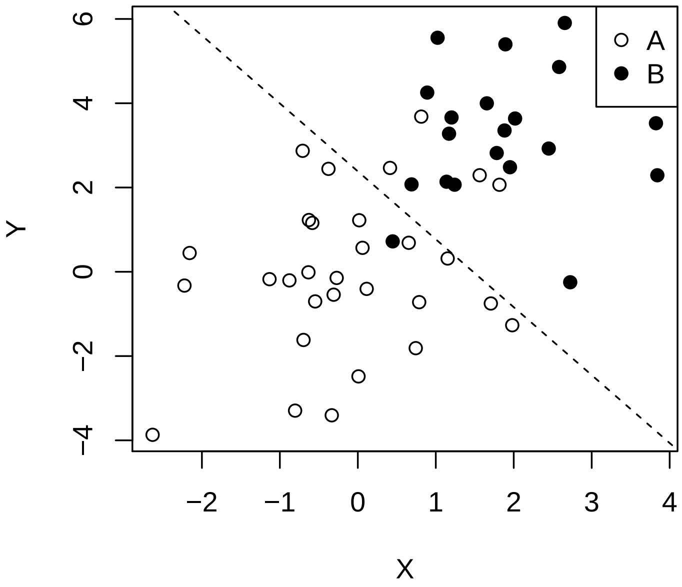
Linear discriminant analysis

- A predictive model that finds the best *linear* separation between two classes.
- It is a form of *supervised learning* and is an alternative to logistic regression.
- `lda()` of `MASS`.

LDA Example

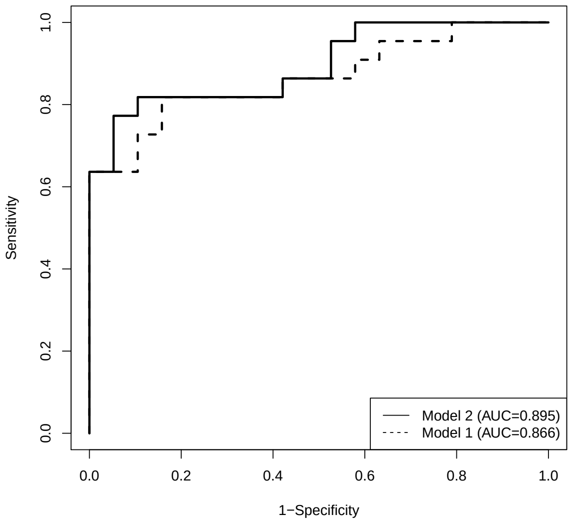
```
## Group A.
A <- data.frame(X=rnorm(30, 0, 1), Y=rnorm(30, 0, 2))
B <- data.frame(X=rnorm(20, 2, 1), Y=rnorm(20, 3, 2))
mydata2 <- cbind(rbind(A, B),
                 Grp=c(rep("A", 30), rep("B", 20)))
library(MASS)
## attach() makes objects in a data.frame visible
### at the top-level
rm(list=c("X","Y","Grp")); attach(mydata2)
mod7 <- lda(Grp~X+Y)
ss1 <- mod7$scaling              # discriminant function coefs
ss1
cc1 <- mean(ss1[1] * X + ss1[2] * Y) #cutoff point
cc1
detach(mydata2)
```

```
with(mydata2, plot(X, Y, pch=ifelse(Grp=="A", 1, 19)))
abline(cc1/ss1[2], -ss1[1]/ss1[2], lty=2)
legend("topright",
       legend=c("A", "B"),
       pch=c(1,19))
```



ROC Curves

- Receiver operating characteristic (ROC) curve. Package: `ROCR` package.
- Trade-off between type I and type II errors.



Survival Analysis

- Assume that we want to establish the association between some clinical covariates and the *survival time* of patients.
- What if many subjects survived the trial?
- Censored data shouldn't be treated as "missing" or "truncated".
- One of the most popular model: Cox proportional hazards model. Function: `coxph()` from the `survival` package.

Cluster analysis

- Hierarchical cluster analysis.
- Distance-based methods. `kmeans()`, `skmeans()`.
- Model-based approaches. Package`Mclust` includes several most popular models.
- Specialized methods. Time course data etc.

Power analysis

- What you need: A comparable study from which you can find: $n_1, n_2, d = \dfrac{|\mu_1 - \mu_2|}{\sigma_{\text{pool}}}$.
- Justify that the proposed study is comparable to that prior study.
- Package: `pwr`, `pwr.t2n.test()`, `pwr.anova.test()` etc.

Other topics

- Financial engineering, time series data analysis, machine learning, Bayesian inference, image analysis, high-performance computing, etc.
- Differential equation solvers, inverse problem (will be discussed in Profs. Wu and Miao's class).
- Bioinformatics related topics will be discussed in detail on 6/3.

Bibliography