Statistics Introduction

Christof Angermueller

October 20, 2014

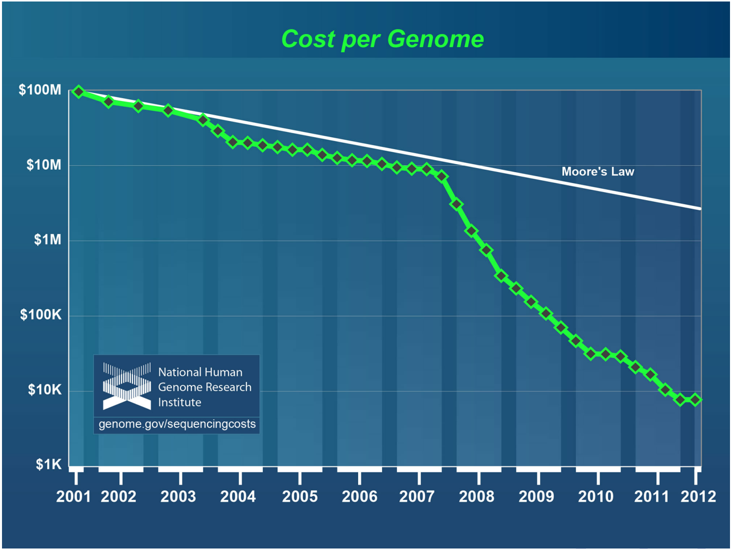
1. Introduction
2. Descriptive statistics
3. Hypothesis testing
4. PCA

About me

- 2nd year PhD student EMBL-EBI
- Supervisor: Oliver Stegle
- Machine learning, Statistics, Biological data analysis
- @cangermueller
- http://cangermueller.com

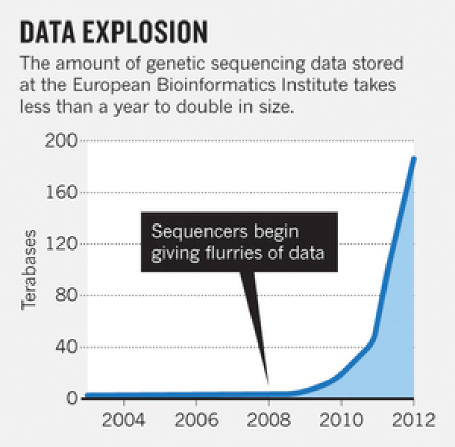
Sequencing costs

- Rapidly declining sequencing costs per genome



Big data in biology

- EBI stores 10 peta-bytes: 10 000 000 000 000 000 bytes
- 2 peta-beta bytes of genomic data
- Huge computational challenges
- Huge biological potential



Why should I care about statistics?

- Because you can analyse your own data
- Because it allows you to tap the reservoir of biological data
- Because it increases your chance to find a job after your PhD
- Because it is fun!

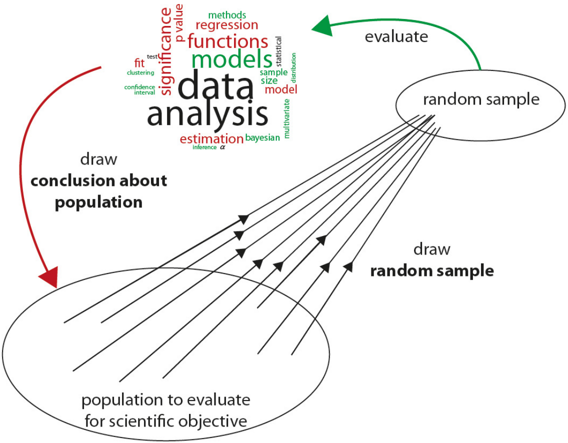
1 Introduction

2 Descriptive statistics

3 Hypothesis testing

4 PCA

Statistical inference



Type-1 Diabetes (T1D) data

- 9627 samples from human population
- Classified as diabetic or non-diabetic
- Genotyped for T1D risk genes

```r
xtable(t1d[sample(nrow(t1d))[1:10],])
```

	t1d	hlacat	sex	age	european	ptpn22	il10	ctla4	bach2	erbb3	gab3
5579	yes	3	f	14.26	TRUE	0	2	1	2	1	0
8872	no	1	f	9.46	TRUE	0	2	2	2	0	0
9254	no	1	f	5.41	TRUE	1	2	1	1	0	1
193	yes	6	f	14.07	TRUE	0	2	1	1	1	0
3279	yes	4	m	2.25	TRUE	1	2	2	2	1	0
909	yes	6	m	9.57	TRUE	1	0	2	1	0	0
3486	yes	4	f	13.78	TRUE	0	2	0	1	0	0
1810	yes	6	f	4.50	TRUE	0	1	2	1	1	0
4251	yes	3	f	4.50	TRUE	1	2	2	1	1	1
9126	no	1	m	9.57	TRUE	0	2	1	1	0	0

T1D variables

- Y: **Output variable, target variable**
  - T1D yes or no
- X_i: **Input variables, explanatory variables, covariates**
  - Sex
  - Age
  - European
  - PTPN22, IL10, CTLA4, ...

Types of variables

Discrete

- $X \in \{s_1, s_2, \dots, s_n\}$
- **Binary**
  european $\in$ {TRUE, FALSE}
- **Categorical**
  sex $\in$ {f, m}
- **Ordinal**
  age $\in$ {child, adult, elder}
- **Integer**
  ptpn22 $\in \{0, 1, 2\}$

Continuous

- $X \in \mathbb{R}$
- age $\in [0, 0.01, \dots, \mathrm{inf}[$

Binary < Categorical < Ordinal < Integer < Continuous

Operations on variables

Count	Binary, categorical, ordinal, integer, continuous
Median	Ordinal, integer, continuous
Median	Ordinal, integer, continuous
Mean	Integer, continuous

Count

Definition (Count)

- How often does x appear?
- Binary, categorical, ordinal, integer, continuous

```
table(t1d$t1d)


##
##   no  yes
## 1766 7861


table(t1d$sex)


##
##    f    m
## 4721 4906
```

Median

Definition (Median)

- Value in the middle
- Ordinal, integer, continuous

```
median(t1d$hlacat)

## [1] 3

median(t1d$age)

## [1] 7.99
```

Quantile

Definition (Quantile)

- $q_p(X)$ is value x, s.t. $q\%$ of all $y \in Y$ are smaller than x
- $q_{0.5}(X) = \text{median}(X)$
- Ordinal, integer, continuous

```
quantile(t1d$hlacat)


##   0%  25%  50%  75% 100%
##    1    2    3    6    6


quantile(t1d$age)


##     0%    25%    50%    75%   100%
##  0.000  5.234  7.990 10.674 22.138
```

Mean

Definition (Mean)

- $\text{mean}(X) = \frac{1}{|X|} \sum_{x \in X} x$
- Integer, continuous variables

```
mean(t1d$age)


## [1] 7.976
```

Random variable

- A random variable X has a random outcome $x \in \mathcal{D}$
- $\mathcal{D}$ is the **domain** of X
- **Discrete** X: $\mathcal{D} \subseteq \mathbb{Z}$, e.g. $\mathbb{D} = \{0, 1, 2, \dots\}$
- **Continuous** X: $\mathcal{D} \subseteq \mathbb{R}$, e.g. $\mathbb{D} = [0.0, 0.1, 0.2, \dots [$
- $P(X = x)$ is the **probability** that X has outcome x
- $E[X] = \sum_{x \in \mathcal{D}} P(X = x)x$ is the **expected value** of X
- $Var[X] = \sum_{x \in \mathcal{D}} P(X = x)(x - E[X])^2$ is the **variance** of X
- $Sd[X] = \sqrt{Var[X]}$ is the **standard deviation** of X

Discrete Random Variable

- $f(x) = P(X = x)$ is the **Probability Mass Function (PMF)** of X
- $F(X) = P(X \leq x)$ is the **Cumulative Distribution Function (CDF)** of X

Examples

- Bernoulli(p): $\mathcal{D} = \{0, 1\}$
- Binomial(n, p): $\mathcal{D} = \{0, 1, \ldots, n\}$
- Poisson(λ): $\mathcal{D} = \{0, 1, \ldots, +\inf\}$

Bernoulli distribution

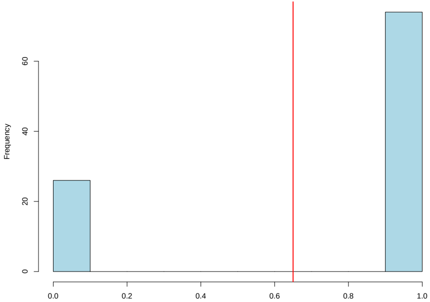
- The gender X with $\mathcal{D} = \{f, m\}$ can be modelled as Bernoulli distributed:

```r
p <- 0.65
exp <- p
samples <- rbinom(100, 1, p)
hist(samples, col='lightblue', main='', xlab='')
abline(v=exp, col='red', lwd=2)
```

$$X \sim \text{Bernoulli}(p)$$

$$E[X] = p$$

$$Var[X] = p(1-p)$$



Bernoulli distribution

- What is the rate p of females of T1D samples?
- The **maximum likelihood** estimator $\hat{p}$ of p is:

$$\hat{p} = \frac{1}{n} \sum_{i} x_i$$

```
p <- sum(t1d$sex == 'f') / nrow(t1d)
p

## [1] 0.4904
```

Binomial distribution

- The number of females Y in a new cohort of $n = 50$ people is binomial distributed:

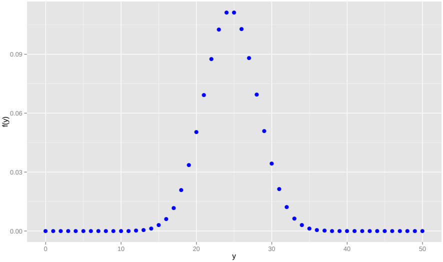
$$Y \sim \text{Binomial}(n, p)$$

- Assume that $p = \hat{p} = 0.4904$ is the same as before ...

Binomial distribution

- then the probability $f(y)$ to observe k females is:

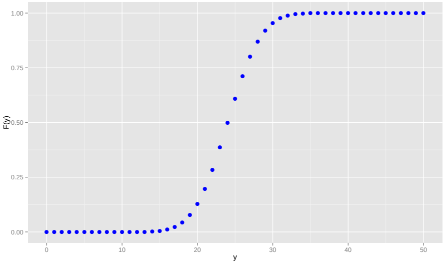
```
n <- 50
y <- seq(0, n)
f_y <- dbinom(y, n, p)
d <- data.frame(x=y, y=f_y)
ggplot(d, aes(x=x, y=y)) + geom_point(color='blue', size=3) + xlab('y') + ylab('f(y)')
```



Binomial distribution

- and the probability $F(Y)$ to observe up to k females:

```
n <- 50
y <- seq(0, n)
F_y <- pbinom(y, n, p)
d <- data.frame(x=y, y=F_y)
ggplot(d, aes(x=x, y=y)) + geom_point(color='blue', size=3) + xlab('y') + ylab('F(y)')
```



Continuous Random Variable

- $f(x) = P(X \in]x-\epsilon; x+\epsilon[)$ is the **Probability Density Function (PDF)** of X
- $F(X) = P(X \leq x)$ is the **Cumulative Distribution Function (CDF)** of X

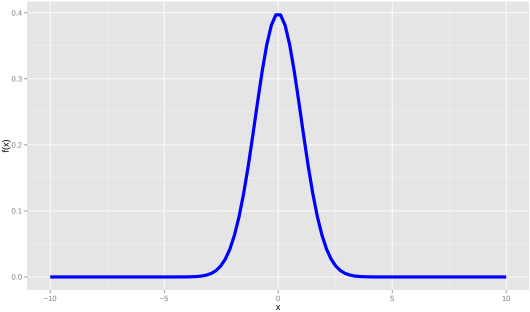
Examples

- Normal(μ, σ^2): $\mathcal{D} = \mathbb{R}$
- Exponential(λ): $\mathcal{D} = \mathbb{R}^+$
- Beta(a, b): $\mathcal{D} = [0, \ldots, 1]$

Normal distribution

- $X \sim N(\mu, \sigma^2)$
- $E[X] = \mu,\ Var[X] = \sigma^2$
- $Z \sim N(0, 1)$ is **standard normal** distribution
- $X \sim N(\mu, \sigma^2) \rightarrow \frac{X-\mu}{\sigma} \sim N(0, 1)$

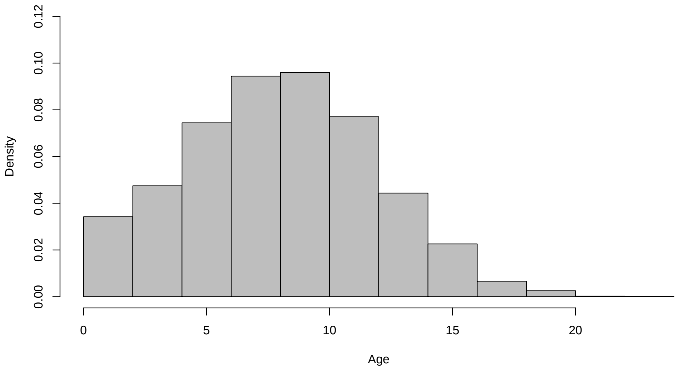
```
x <- seq(-10, 10, len=100)
y <- dnorm(x, mean=0, sd=1.0)
d <- data.frame(x=x, y=y)
ggplot(d, aes(x=x, y=y)) + geom_line(color='blue', size=2) + xlab('x') + ylab('f(x)')
```



Example

- How is the **age** distributed in the T1D cohort?

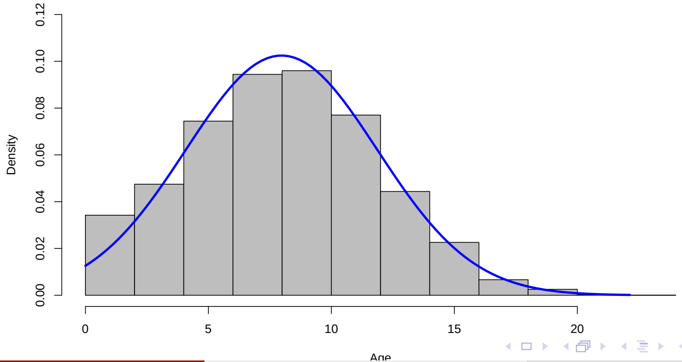
```
hist(t1d$age, col='grey', xlab='Age', ylab='Density', prob=TRUE, main=NULL, ylim=c(0, 0.12))
```



Example

- Estimate μ and σ via maximum likelihood:

```
mu <- mean(t1d$age)
sigma <- sd(t1d$age)
x <- seq(min(t1d$age), max(t1d$age), len=100)
y <- dnorm(x, mean=mu, sd=sigma)
hist(t1d$age, col='grey', xlab='Age', ylab='Density', prob=TRUE, main=NULL, ylim=c(0, 0.12))
lines(x, y, col='blue', lwd=3)
```



1 Introduction

2 Descriptive statistics

3 Hypothesis testing

4 PCA

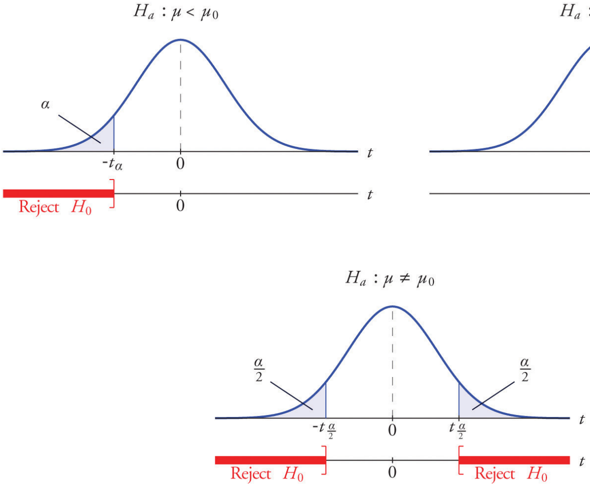
Relationship between variables

- How are X and Y related?
- X: input variable, explanatory variable, covariates
- Y: output variable, target variable
- Is the relationship significant?

Process

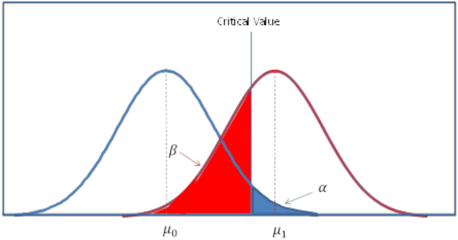
1. Define hypothesis
   - One-sided: $\mu < \mu_0$, $\mu > \mu_0$
   - Two-sides: $\mu = \mu_0$
2. Defined region of rejection $\mathcal{R}_\alpha$ depending on significance level α
3. Collect data $\mathcal{D}$
4. Compute test statistic $T_\mathcal{D}$ for data $\mathcal{D}$
5. Reject H_0 if $T_\mathcal{D} \notin R_\alpha$

Process


$H_a : \mu < \mu_0$
α
$-t_\alpha$
0
t
Reject H_0
0
t
$H_a : \mu > \mu_0$
α
0
t_α
t
0
Reject H_0
t
$H_a : \mu \neq \mu_0$
$\frac{\alpha}{2}$
$\frac{\alpha}{2}$
$-t_{\frac{\alpha}{2}}$
0
$t_{\frac{\alpha}{2}}$
t
Reject H_0
0
Reject H_0
t

Type-1 and Type-2 error

	No reject	**Reject**
H_0 **true**	True Negative (TN)	False Positive (FP) Type-1 error, α
H_0 **false**	False Negative (FN) Type-2 error, β	True Positive (TP)



Overview hypothesis tests

	Discrete	**Continuous**
Discrete	Score test Fisher's exact test χ^2 test Logistic regression	Z test Student's t test
Continuous	*Descretize*	Correlation Linear regression

p-value

What is a p-value?

p-value

Definition (p-value)

Probability to observe by chance a test statistic T' that is at least as extreme as the observed test statistic $T_{\mathcal{D}}$, given that H_0 is true.

Testing two proportions

- Are females more likely to develop T1D than males?

$$X_f \sim \text{Bernoulli}(p_f) \quad X_m \sim \text{Bernoulli}(p_m)$$

$$H0 : p_f = p_m$$

```
sex_t1d <- table(t1d$sex, relevel(t1d$t1d, 'yes'))
xtable(sex_t1d)
```

	yes	no
f	3775	946
m	4086	820

Score test

```
prop.test(sex_t1d, conf.level=0.95, alternative='two.sided')


##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  sex_t1d
## X-squared = 17.52, df = 1, p-value = 2.837e-05
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04892 -0.01756
## sample estimates:
## prop 1 prop 2
## 0.7996 0.8329
```

- Depends on central limit theorem
- Requires many samples

Fisher's exact test

```
fisher.test(sex_t1d, conf.level=0.95, alternative='two.sided')


##
##  Fisher's Exact Test for Count Data
##
## data:  sex_t1d
## p-value = 2.789e-05
## alternative hypothesis: true odds ratio is not equal to 1
## 95 percent confidence interval:
##  0.7211 0.8893
## sample estimates:
## odds ratio
##     0.8009
```

- Does not depend on central limit theorem
- Does not require many samples
- Exact test: guarantees false positive rate
- Sometime too conservative

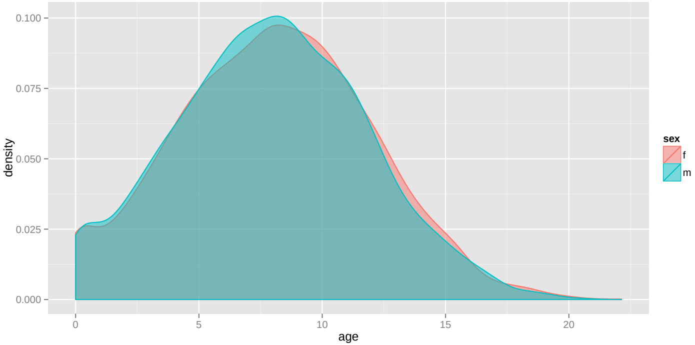
Comparing means

- Are females significantly older than males?

$$X_f \sim \mathcal{N}(\mu_f, \sigma^2) \quad X_m \sim \mathcal{N}(\mu_m, \sigma^2)$$

$$H0 : \mu_f < \mu_m$$

```
ggplot(t1d) + geom_density(aes(x=age, y=..density.., group=sex, fill=sex, color=sex), alpha=.5)
```



Student’s t test

```r
t.test(t1d$age[t1d$sex == 'f'], t1d$age[t1d$sex == 'm'], alternative='greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  t1d$age[t1d$sex == "f"] and t1d$age[t1d$sex == "m"]
## t = 1.644, df = 9594, p-value = 0.0501
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  -7.71e-05       Inf
## sample estimates:
## mean of x mean of y
##     8.042     7.912
```

Testing dependency between variables

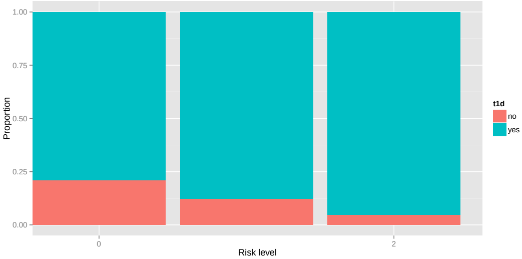
- Does PTPN22 influence the risk of T1D?

$$X_{\text{PTPN22}} \sim \text{Cat}(\{0, 1, 2\}) \quad X_{\text{T1D}} \sim \text{Bernoulli}(p)$$

$$H0 : X_{\text{PTPN22}} \text{ independent of } X_{\text{T1D}}$$

Testing dependency between variables

```
table(t1d$ptpn22, t1d$t1d)


##
##       no  yes
##   0 1441 5381
##   1  313 2239
##   2   12  241


ggplot(t1d) + geom_bar(aes(x=ptpn22, fill=t1d), position='fill') +
scale_x_discrete(limits=c(0, 2)) + xlab('Risk level') + ylab('Proportion')
```



Testing dependency between variables

- Fitting logistic regression model

```
model <- glm(t1d ~ factor(ptpn22), data=t1d, family=binomial)
xtable(model)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.3175	0.0297	44.42	0.0000
factor(ptpn22)1	0.6500	0.0672	9.67	0.0000
factor(ptpn22)2	1.6824	0.2972	5.66	0.0000

Testing dependency between variables

- Likelihood Ratio Test

```
anova(model, test='LRT')


## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: t1d
##
## Terms added sequentially (first to last)
##
##
##                 Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                            9626       9176
## factor(ptpn22)  2      145      9624       9031   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Testing dependency between variables

- Does PTPN22 influence the risk of T1D, accounting for all other variables?

$$X_{\mathrm{PTPN22}} \sim \mathrm{Cat}(\{0,1,2\}) \quad X_{\mathrm{T1D}} \sim \mathrm{Bernoulli}(p)$$

$H0 : X_{\mathrm{PTPN22}}$ independent of X_{T1D}, accounting for $X_{\mathrm{age}}, X_{\mathrm{ERBB3}}, X_{\mathrm{IL10}}, \ldots$

Testing dependency between variables

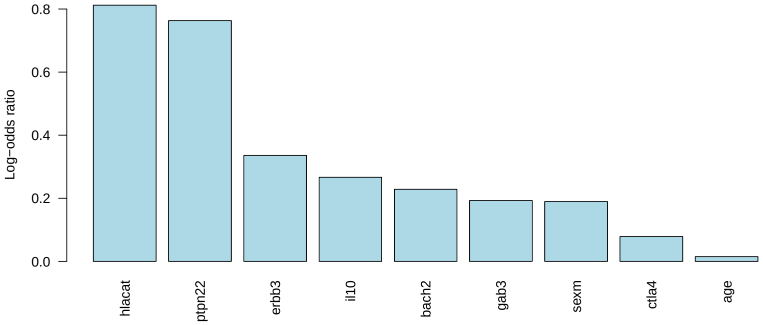
- Fitting logistic regression model

```
model <- glm(t1d ~ .-(european), data=t1d, family=binomial)
xtable(model)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.2612	0.1583	-14.29	0.0000
hlacat	0.8121	0.0240	33.89	0.0000
sexm	0.1896	0.0601	3.16	0.0016
age	0.0151	0.0077	1.95	0.0512
ptpn22	0.7633	0.0673	11.34	0.0000
il10	0.2665	0.0585	4.56	0.0000
ctla4	0.0788	0.0431	1.83	0.0675
bach2	0.2285	0.0422	5.42	0.0000
erbb3	0.3360	0.0447	7.52	0.0000
gab3	0.1930	0.0374	5.16	0.0000

Testing dependency between variables

```
lor <- sort(coef(model)[-1], decreasing=TRUE)
barplot(lor, las=2, col='lightblue', ylab='Log-odds ratio')
```



1 Introduction

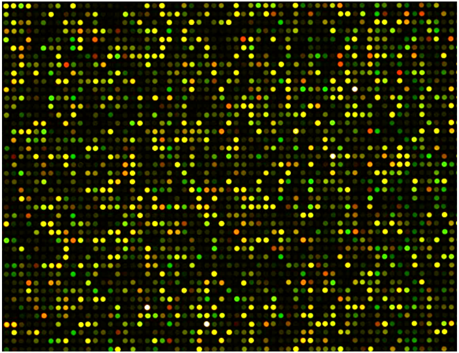
2 Descriptive statistics

3 Hypothesis testing

4 PCA

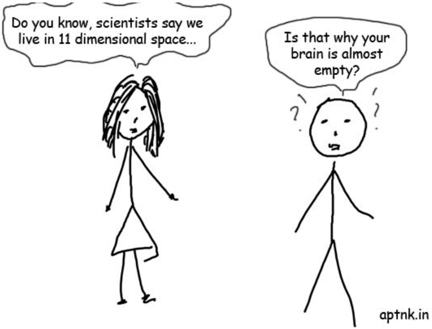
High-dimensional data

- Small n large p
- Few samples n (≈ 100)
- Many features p
  - > 10000 genes
  - > 27000000 CpG sites



Problems

- High storage costs (memory)
- High computational costs (time)
- Visualization?
- Curse of dimensionality

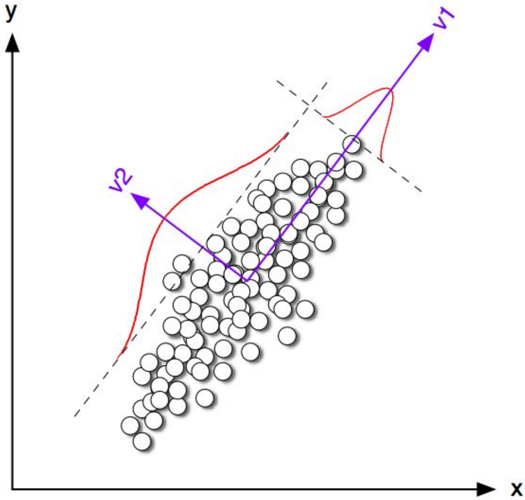

Do you know, scientists say we live in 11 dimensional space...
Is that why your brain is almost empty?
aptnk.in

Principle Component Analysis

- Dimensionality reduction
- Visualization
- Missing values imputation
- Latent factors estimation:
  - Population structure
  - Batch-effects
  - Cell-cycle

Principle components

- Minimize projection error
- Maximize variance



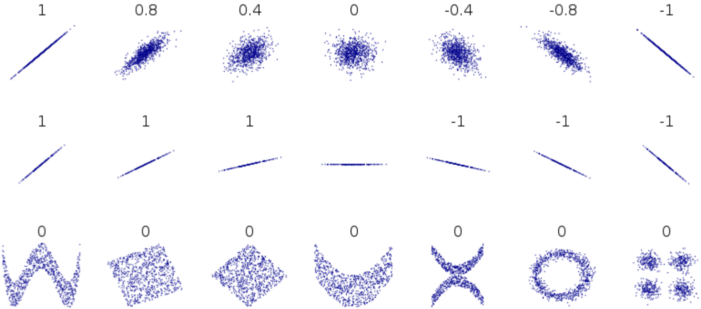
Pearson correlation coefficient

- Measures **linear** dependency between x and y
- $cor(x, y) = \in [-1, +1]$
- $cor(x, y) = 0$: no correlation
- $cor(x, y) = -1$: negative correlation
- $cor(x, y) = +1$: positive correlation

Definition (Pearson correlation coefficient)

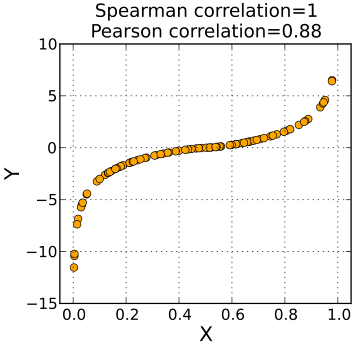
$$\mathrm{cor}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}}$$

Pearson correlation coefficient



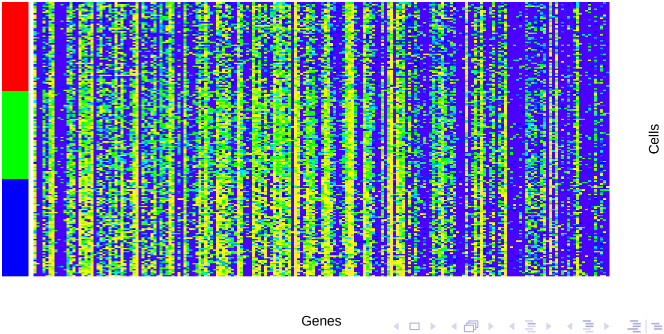
Spearman correlation coefficient

- Measures **monotonic** dependency between x and y
- Pearson correlation coefficient on rank of variables
- $cor(x, y) = \in [-1, +1]$



Embryonic Stem Cell (ESC) data

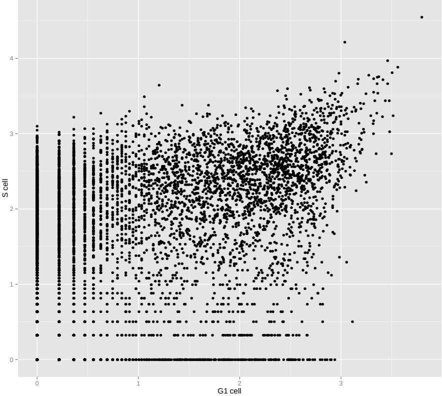
- Single-cell RNA-seq
- $n = 182$ Embryonic stem cells
- $p = 9571$ Genes
- Cell-cycle via Hoechst staining



Correlation between two ESC

```
c1 <- esc$counts[which(esc$cycle == 'G1')[1],]
c2 <- esc$counts[which(esc$cycle == 'S')[1],]
```

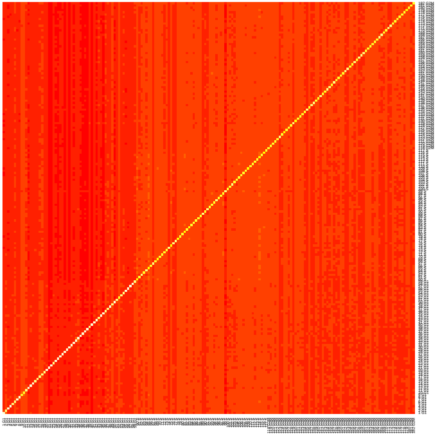
```
qplot(c1, c2, xlab='G1 cell', ylab='S cell')
```



```
cor(c1, c2,
    method='pearson')

## [1] 0.5475

cor(c1, c2,
    method='spearman')

## [1] 0.5786
```

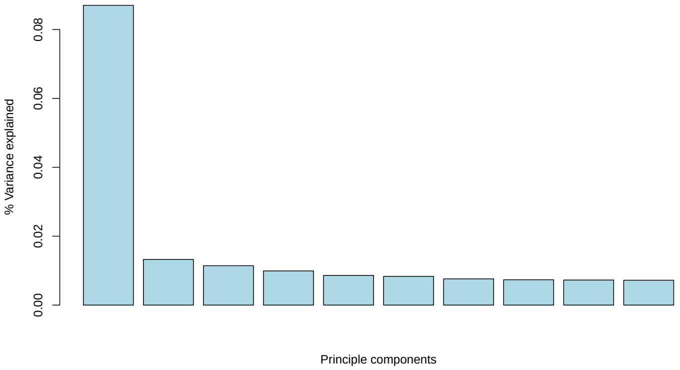
Correlation matrix

```
cor_cells <- cor(t(esc$counts), method='pearson')
heatmap(cor_cells, Rowv=NA, Colv=NA)
```



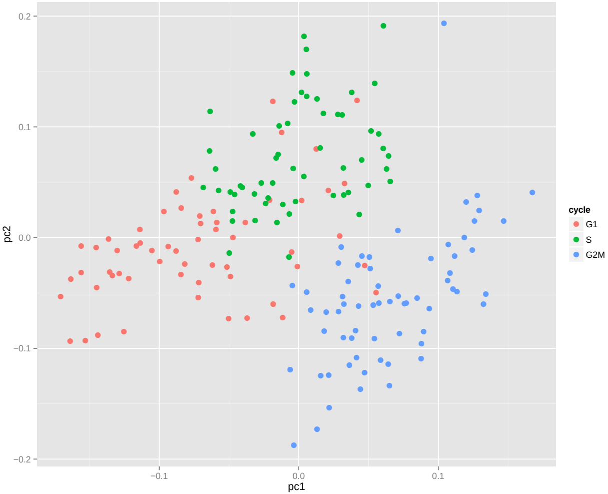
PCA

```
svd <- svd(scale(esc$counts))
dsvd <- as.data.frame(svd$u[, 1:10])
colnames(dsvd) <- paste0('pc', seq(1, ncol(dsvd)))
dsvd$cycle <- esc$cycle
ve <- svd$d^2 / sum(svd$d^2)
barplot(ve[1:10], xlab='Principle components', ylab='% Variance explained',
        col='lightblue')
```



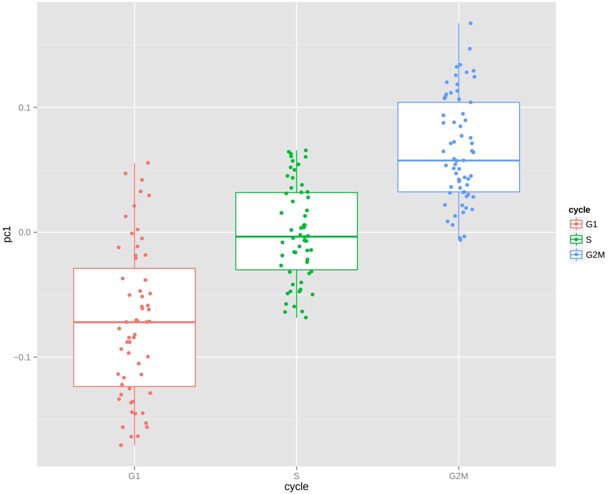
PC1 versus PC2

```
ggplot(dsvd, aes(x=pc1, y=pc2, color=cycle)) + geom_point(size=3)
```



Correlation PC1 and cell-cycle

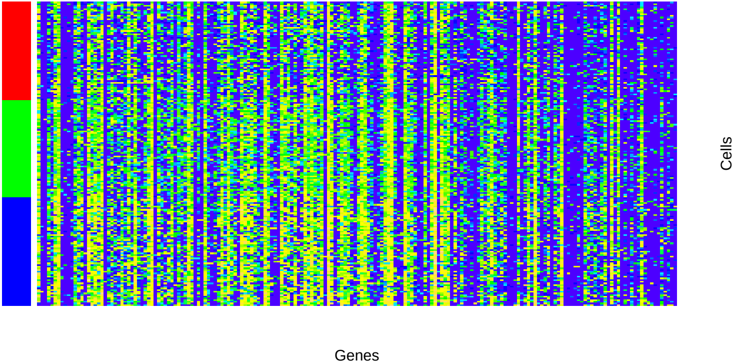
```
ggplot(dsvd, aes(x=cycle, y=pc1, color=cycle)) + geom_boxplot() + geom_jitter(position=position_jitter(width=.1
```



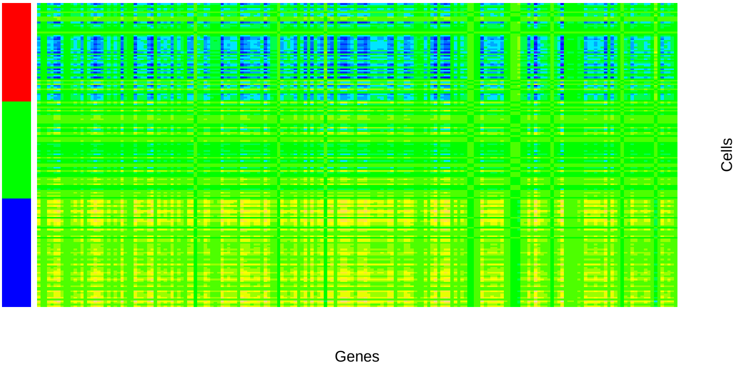
Accounting for cell-cycle

```
svd_compress <- function(svd, k=1) {
  return (svd$u[, 1:k, drop=FALSE] %*% diag(svd$d[1:k], nrow=k) %*% t(svd$v)[1:k,, drop=FALSE])
}
counts_pc1 <- svd_compress(svd, 1)
counts_npc1 <- esc$counts - counts_pc1
```

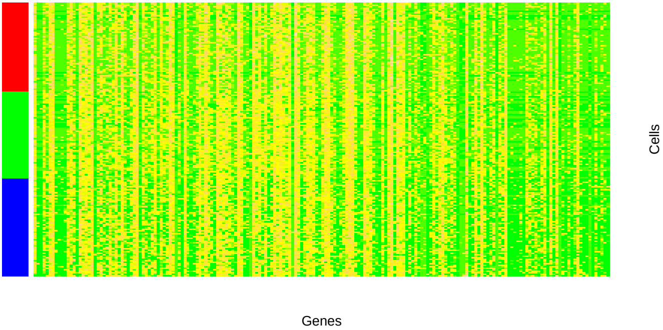
Raw counts



Counts explained by cell-cycle



Counts after adjusting for cell-cycle



Further readings

David J. Balding and Christopher M Bishop.
Handbook of statistical genetics.
Wiley, Chichester [u.a.], 2007.

Coursera.
Data analysis and statistical inference.

Coursera.
Mathematical biostatistics boot camp 1.

Bernard Rosner.
Fundamentals of biostatistics.
Brooks/Cole, Cengage Learning, Boston, 2011.

Cosma Rohilla Shalizi.
Advanced data analysis from an elementary point of view.
Preprint of book found at http://www. stat. cmu. edu/ cshalizi/ADAfaEPoV, 2013.

Questions

Questions?