南方科技大学
SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Machine Learning and NeuroEngineering

机器学习与神经工程

Lecture 8 – Combining data

Quanying Liu （刘泉影）

SUSTech, BME department

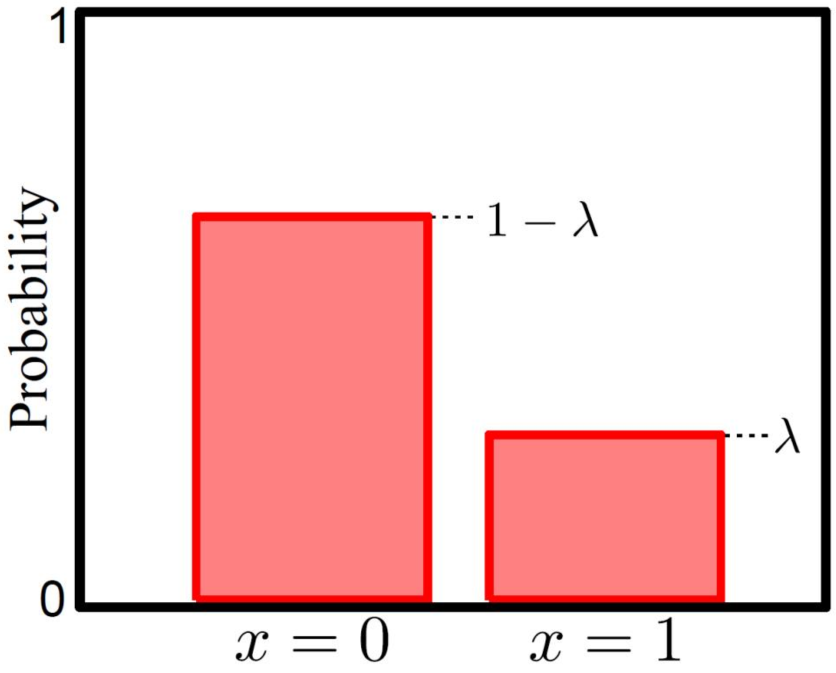
Email: liuqy@sustech.edu.cn

BMEB212, Spring term, 2021

Lecture 6 – Probability Recap

- Random variables
  - Discrete (Bernoulli, Categorical, Binomial, Geometric, Poisson)
  - Continuous (Gaussian, Uniform, Exponential, Beta)
- Joint probability p(x, y)
- Marginalization / Law of Total Probability
- Conditional Probability p(x | y)
- Bayes' Rule
- Expectation & Conditional expectation
- Extension to N random variables

Bernoulli Distribution



$$
\begin{aligned}
Pr(x=0) &= 1-\lambda \\
Pr(x=1) &= \lambda.
\end{aligned}
$$

or

$$Pr(x) = \lambda^{x}(1-\lambda)^{1-x}$$

For short we write:

$$\boxed{p(x) = \mathrm{Ber}(x|\lambda)}$$

Bernoulli distribution describes situation where only two possible outcomes: x = 0 / x = 1 (e.g. failure/success)

Takes a single parameter: $\lambda \in [0, 1]$

Binomial Distribution

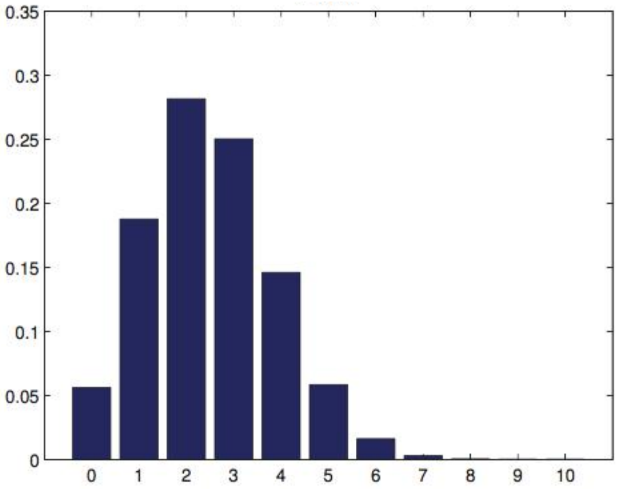
Suppose we toss a coin n times. Let $X \in \{0, \ldots, n\}$ be the number of heads. If the probability of heads is θ, then we say X has a **binomial** distribution, written as $X \sim \text{Bin}(n, \theta)$. The pmf is given by

$$\text{Bin}(k|n,\theta) \triangleq \binom{n}{k} \theta^k (1-\theta)^{n-k}$$

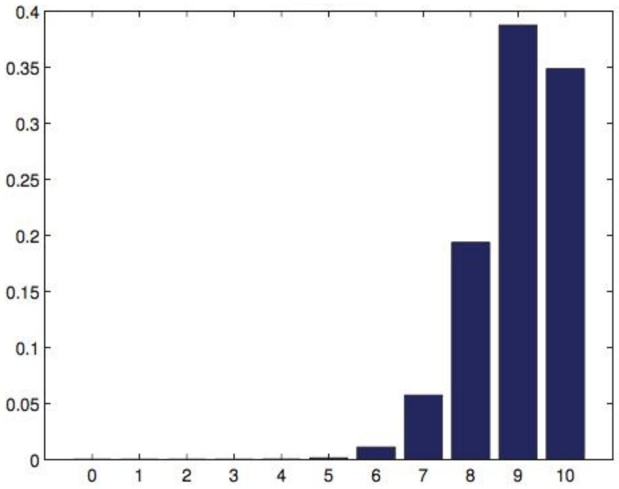
where

$$\binom{n}{k} \triangleq \frac{n!}{(n-k)!k!}$$

n=10, theta = 0.25

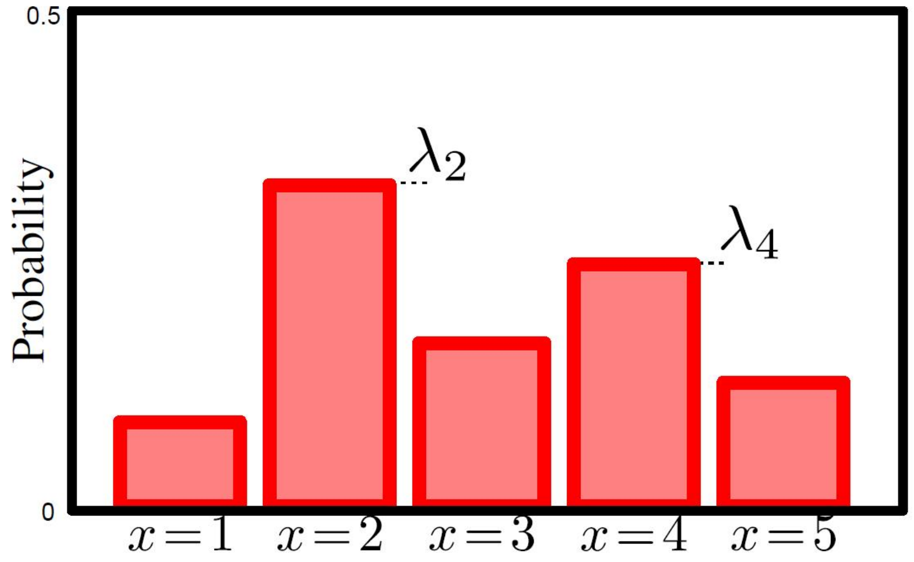


n=10, theta = 0.80



$$\text{mean} = \theta, \quad \text{var} = n\theta(1-\theta)$$

Categorical Distribution



$$Pr\,(x = k) = \lambda_k$$

or can think of data as vector with all elements zero except kth e.g. $\mathbf{e}_4$ = [0,0,0,1,0]

$$Pr\,(x = \mathbf{e}_k) = \prod_{j=1}^{K} \lambda_j^{\mathbf{e}_{kj}} = \lambda_k$$

where $\mathbf{e}_{kj}$ is the j-th element of $\mathbf{e}_k$

For short we write: $\boxed{p(x) = \text{Cat}(x|\lambda)}$

Categorical distribution describes situation where K possible outcomes:

x = 1 , … , x = k , … , x = K.

Takes K parameters $\lambda_k \in [0,1]$ where $\sum_{k=1}^{K} p(X=k) = \sum_{k=1}^{K} \lambda_k = 1$

$\lambda = \{\lambda_1, \cdots, \lambda_K\}$

Poisson Distribution

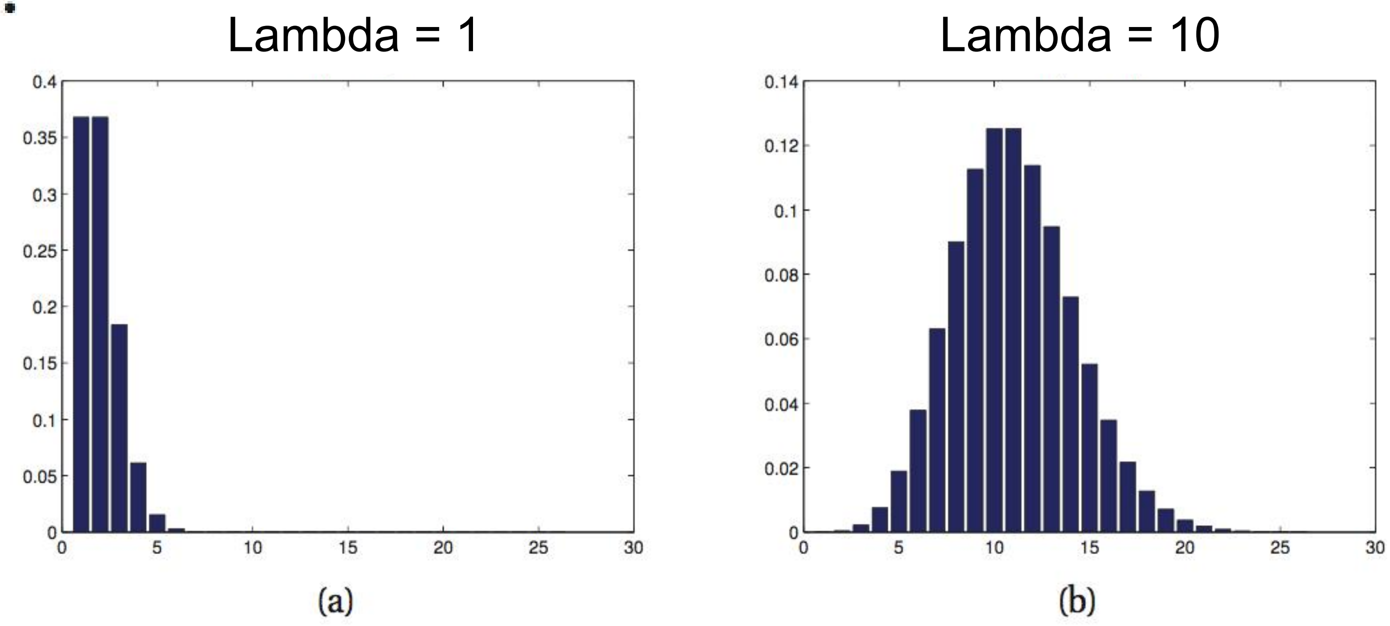
We say that $X \in \{0, 1, 2, \ldots\}$ has a **Poisson** distribution with parameter $\lambda > 0$, written $X \sim \mathrm{Poi}(\lambda)$, if its pmf is

$$\mathrm{Poi}(x|\lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$$

normalization constant
to ensure the distribution sums to 1

The Poisson distribution is often used as a model for **counts of rare events** like radioactive decay and traffic accidents.

In neuroscience, we used Poisson distribution to model **the counts of neuron spikes**.



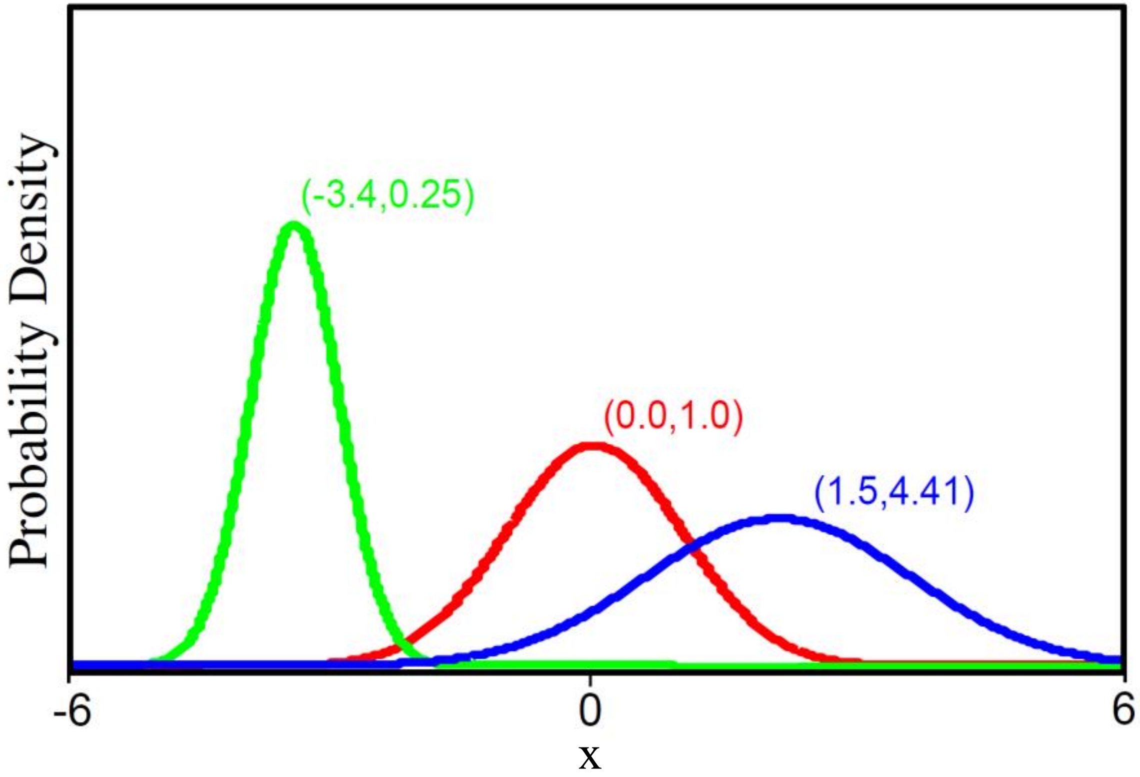
MLaPP, Kevin Murphy

Gaussian / Normal Distribution

$X \sim N(\mu, \sigma^2)$

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\boxed{-\frac{(x-\mu)^2}{2\sigma^2}}}$$

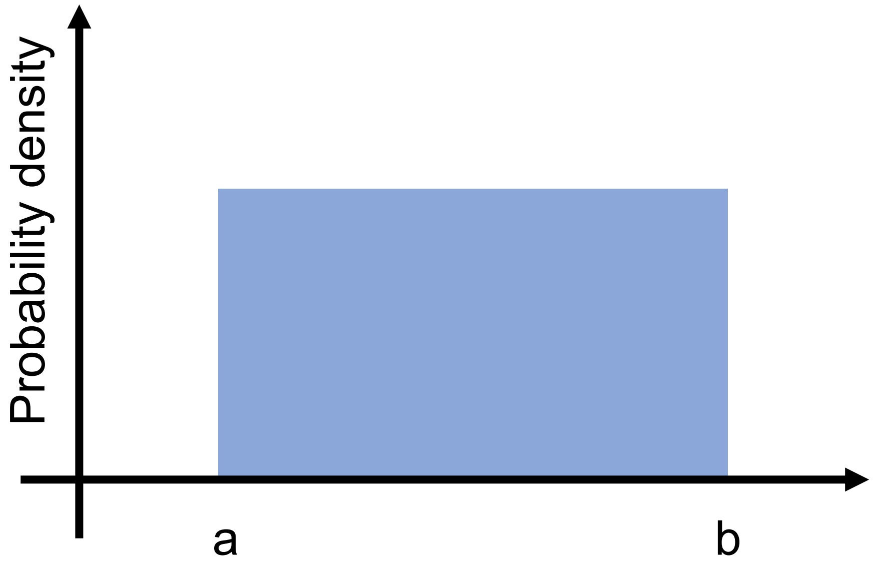
≤ 0



2 parameters:
mean μ and variance $\sigma^2 > 0$

Adapted from S. Prince

Uniform Distribution



$$\mathrm{Unif}(x|a,b) = \frac{1}{b-a}\mathbb{I}(a \le x \le b)$$

Mean = (b+a)/2

var = (b-a)^2/12

We usually use it as an uninformative prior.

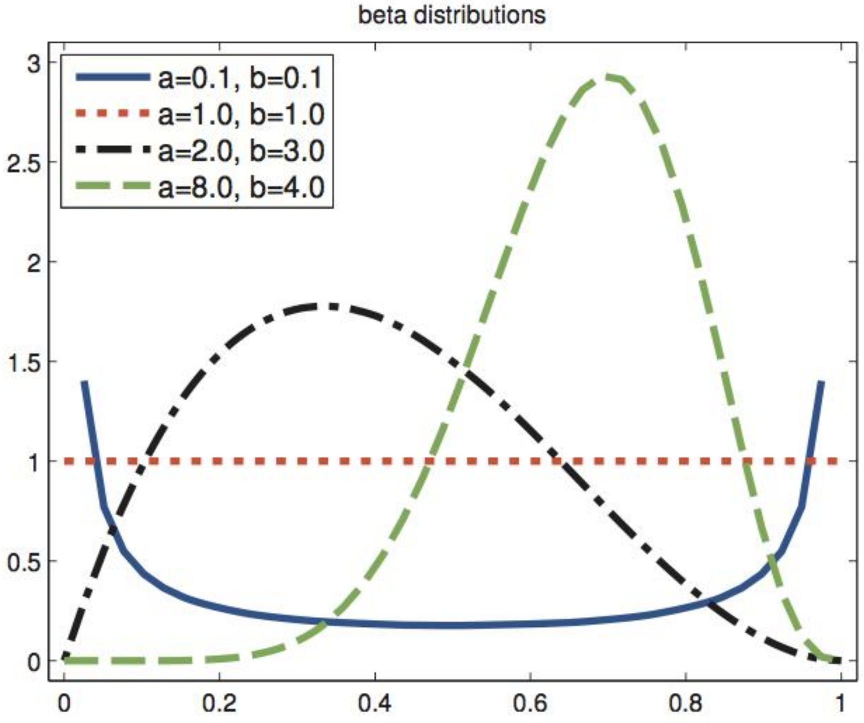
$$\mathrm{var}\left[X\right] \triangleq \mathbb{E}\left[(X-\mu)^2\right] = \int (x-\mu)^2 p(x)dx$$

MLaPP, Kevin Murphy

Beta Distribution

The **beta distribution** has support over the interval $[0, 1]$ and is defined as follows:

$$\mathrm{Beta}(x|a,b) = \frac{1}{B(a,b)} x^{a-1}(1-x)^{b-1} \qquad B(a,b) \triangleq \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$$



$$\mathrm{mean} = \frac{a}{a+b}$$

$$\mathrm{mode} = \frac{a-1}{a+b-2}$$

$$\mathrm{var} = \frac{ab}{(a+b)^2(a+b+1)}$$

Beta distribution is important, for we usually use it as a conjugate prior of Binomial (or Bernoulli) process.

MLaPP, Kevin Murphy

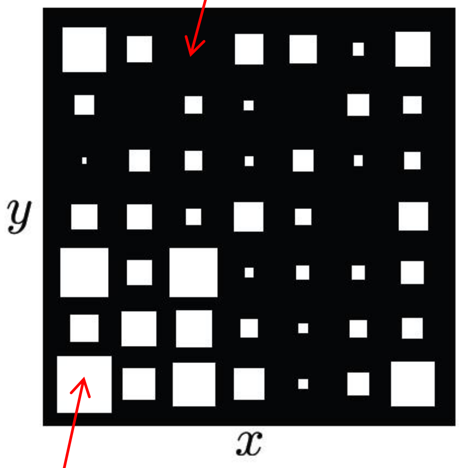
Joint Probability

- If we observe two random variables x & y multiple times, then some combinations of outcomes more likely than others
- This information captured by joint probability distribution
- Written as p(x, y), which is read as “joint probability distribution of x and y”

Adapted from S. Prince

Joint Probability p(x, y)

unlikely

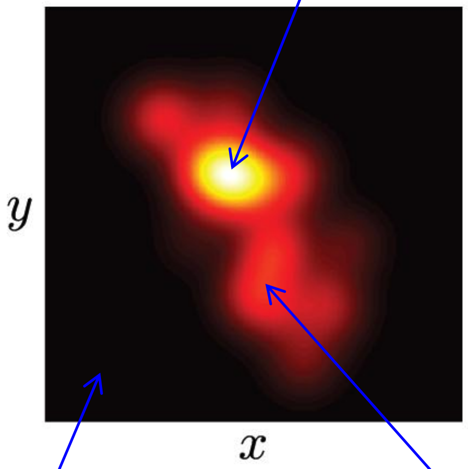


x & y discrete

most likely (x, y) pair

$$\sum_y \sum_x p(x, y) = 1$$

most likely (x, y) pair

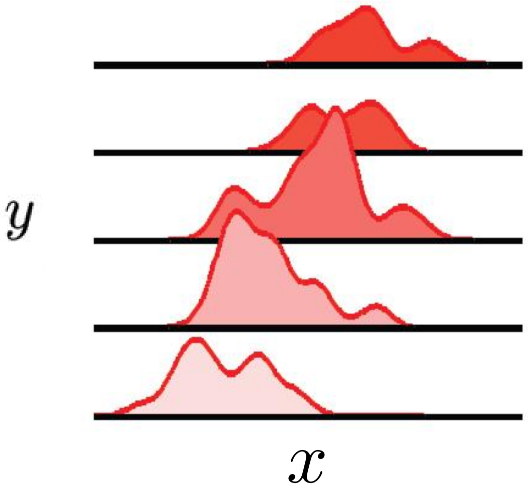


x & y continuous

unlikely

somewhat likely

$$\int_y \int_x p(x, y) dx dy = 1$$



x is continuous,
y is discrete

$$\sum_y \int_x p(x, y) dx = 1$$

Adapted from S. Prince

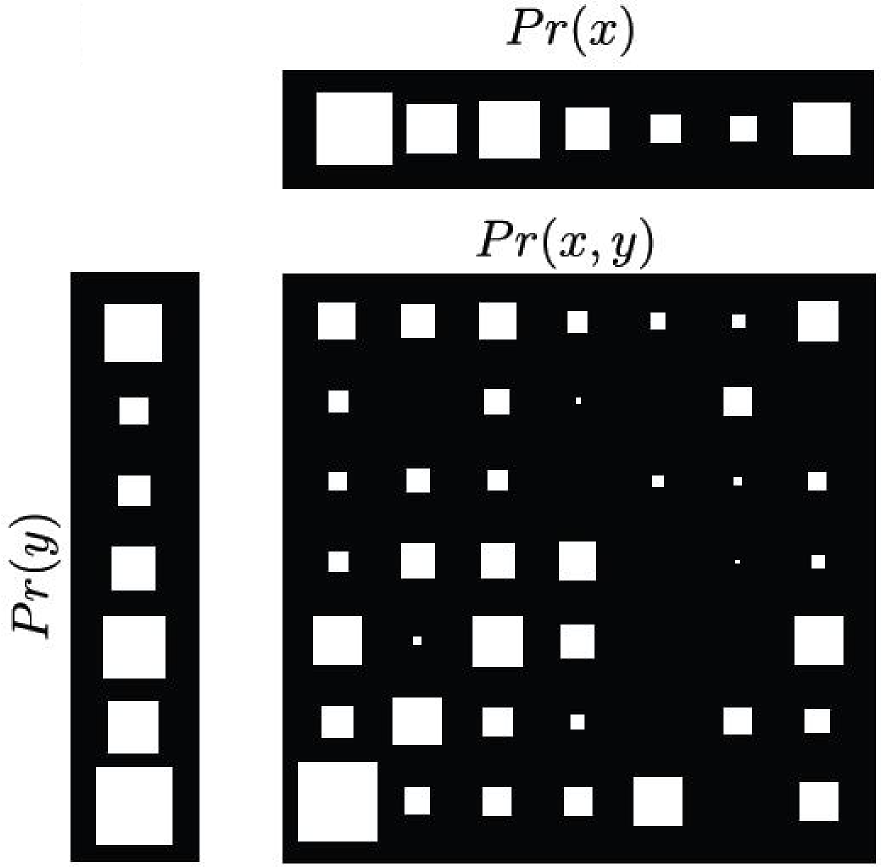
Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s).

This is called marginalization.

$$p(x) = \sum_y p(x,y)$$

$$p(y) = \sum_x p(x,y)$$



Adapted from S. Prince

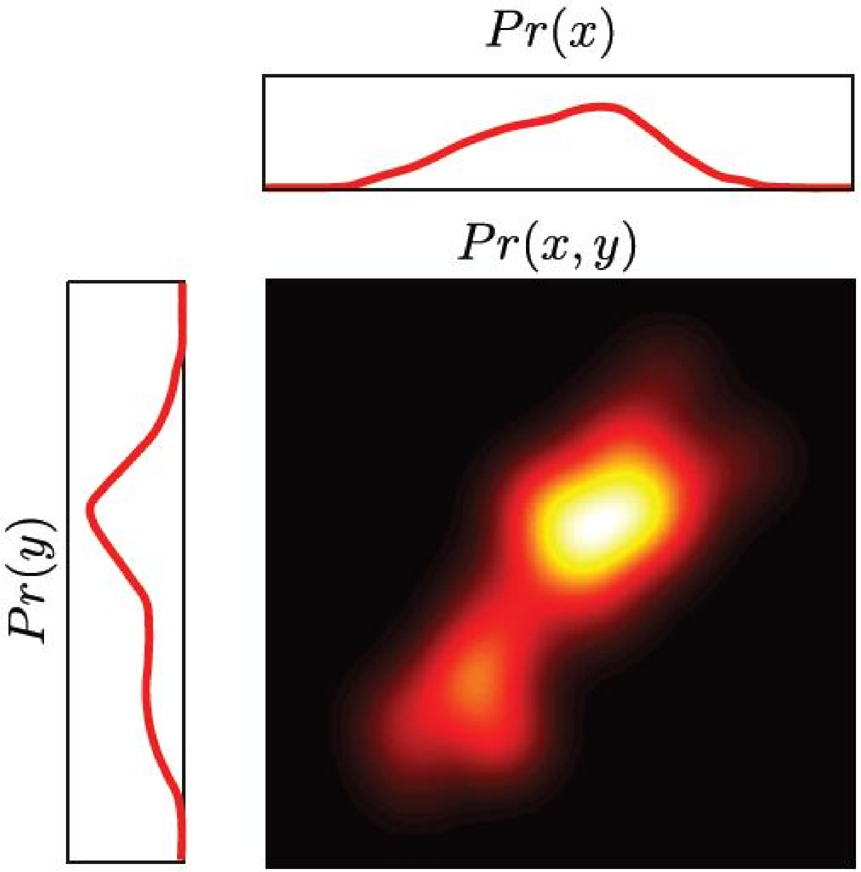
Marginalization / Law of Total Probability

We can recover probability distribution of any variable in a joint distribution by integrating (or summing) over the other variable(s).

This is called marginalization.

$$p(x) = \int_y p(x, y) dy$$

$$p(y) = \int_x p(x, y) dx$$



Adapted from S. Prince

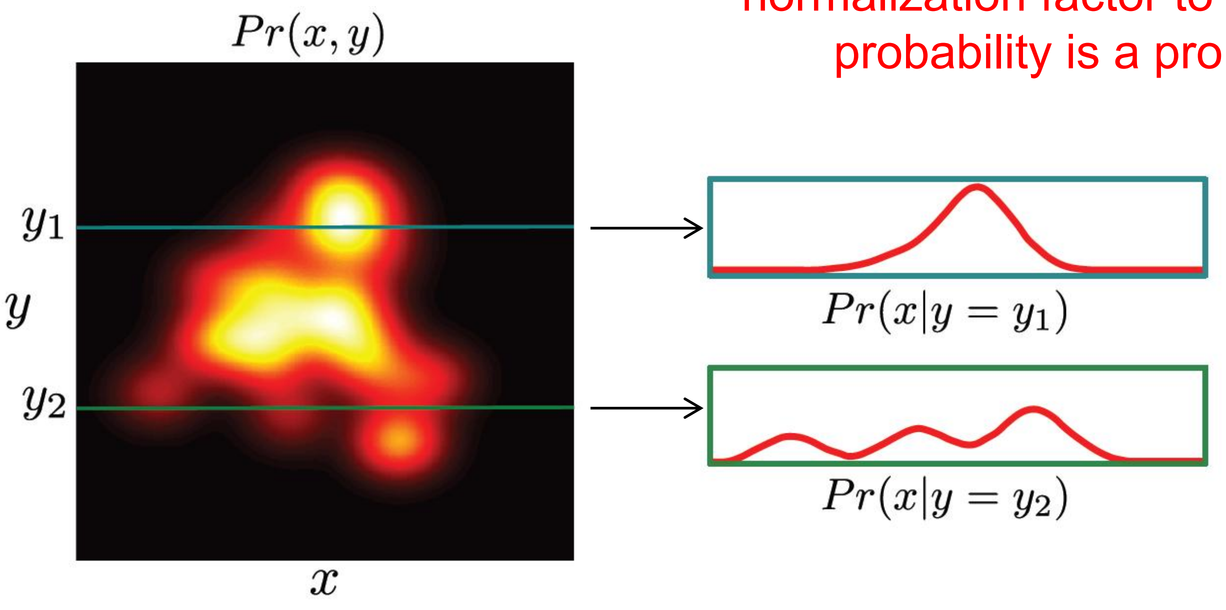
Conditional Probability

- Conditional probability can be computed from joint probability.

slice of joint distribution

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{p(y=y^*)} = \frac{p(x,y=y^*)}{\int p(x,y=y^*)dx}$$

normalization factor to ensure conditional probability is a proper distribution



$$\int_x p(x|y=y_1)dx = 1$$

$$\int_x p(x|y=y_2)dx = 1$$

Adapted from S. Prince

Conditional Probability

$$p(x|y=y^*) = \frac{p(x,y=y^*)}{p(y=y^*)} = \frac{p(x,y=y^*)}{\int p(x,y=y^*)dx}$$

- More usually written in compact form

$$p(x|y) = \frac{p(x,y)}{p(y)}$$

- Can be re-arranged to give

$$p(x,y) = p(y)p(x|y)$$
$$p(x,y) = p(x)p(y|x)$$

Adapted from S. Prince

Deriving Bayes' Rule (y continuous)

From before:

$$p(x,y) = p(y)p(x|y)$$

$$p(x,y) = p(x)p(y|x)$$

Equate RHS

Combining:

$$p(y)p(x|y) = p(x)p(y|x)$$

Re-arranging:

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)}$$

$$p(x) = \sum_{y} p(x,y)$$

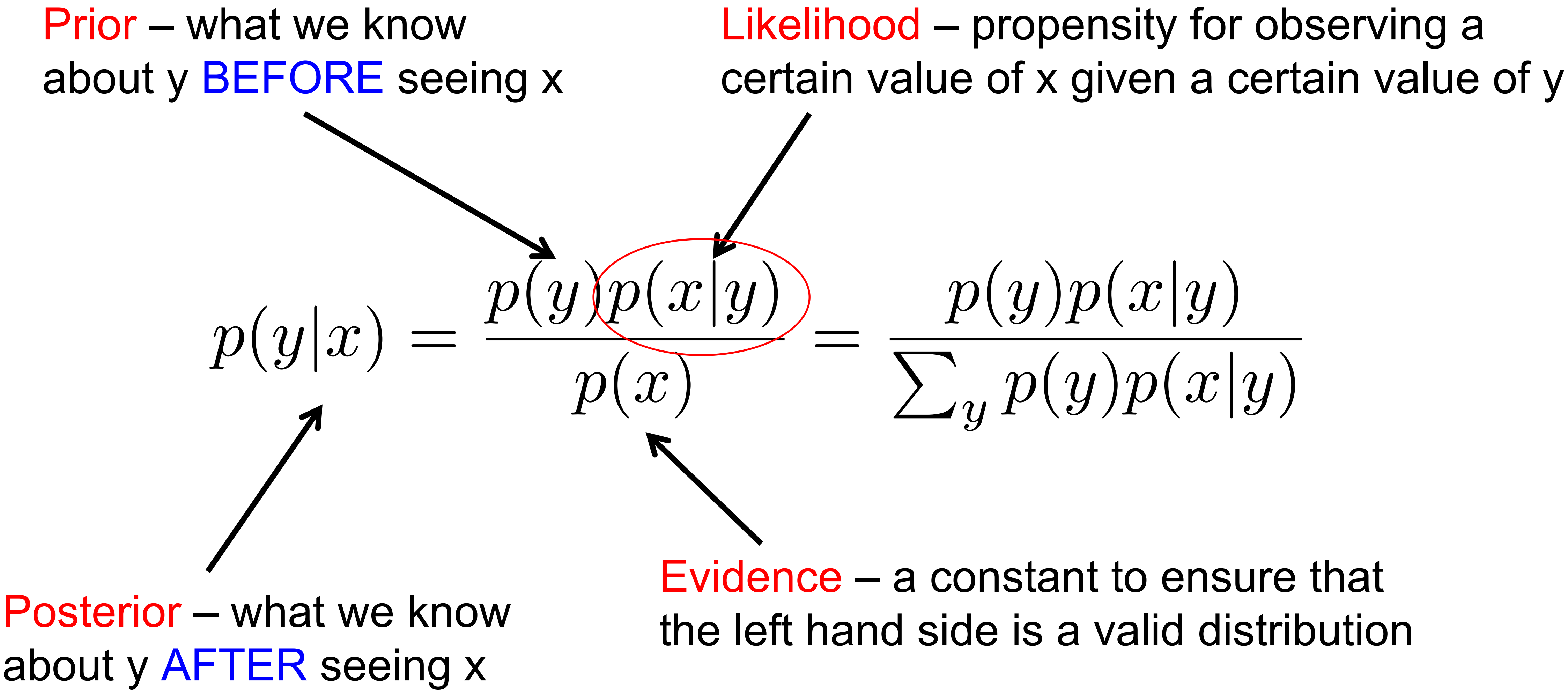
$$= \frac{p(y)p(x|y)}{\sum_y p(x,y)}$$

$$p(x,y) = p(y)p(x|y)$$

$$= \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Adapted from S. Prince

Bayes’ Rule



Prior – what we know about y BEFORE seeing x

Likelihood – propensity for observing a certain value of x given a certain value of y

$$p(y|x) = \frac{p(y)p(x|y)}{p(x)} = \frac{p(y)p(x|y)}{\sum_y p(y)p(x|y)}$$

Posterior – what we know about y AFTER seeing x

Evidence – a constant to ensure that the left hand side is a valid distribution

Adapted from S. Prince

Independence

- If x & y are independent, then knowing x tells us nothing about y (and vice versa):

$$p(x|y) = p(x)$$
$$p(y|x) = p(y)$$

- If x & y are independent, then joint distribution factorizes into product of marginal distributions:

$$\begin{aligned} p(x, y) &= p(x)p(y|x) \\ &= p(x)p(y) \end{aligned}$$

- Conversely, if joint distribution can be factorized into product of marginal distributions, then x & y are independent

Expectation: Mean and Variance

$$E[f(x)] = \int f(x)p(x)dx$$

- If $f(x) = x$
  - $E[f(x)] = E(x) = \mu_x$, the “mean of x”
  - If we observe x many (infinite) times and average, we get μ_x

- If $f(x) = (x - \mu_x)^2$
  - $E[f(x)] = E[(x - \mu_x)^2] = \sigma_x^2$
  - $\sigma_x^2 = Var(x)$ called “variance”; σ_x called ”standard deviation”
  - If we observe x many (infinite) times and average square of difference between each observation and μ_x, we get σ_x^2
  - Measure how likely x is going to be far away from mean

Expectation for X, Y

- Expectation tells us the expected or average value of some function $f(x,y)$ taking into account $p(x,y)$

$$E[f(x,y)] = \int\int f(x,y)p(x,y)dxdy$$

- Special case: $f(x,y) = (x–\mu_x)(y–\mu_y)$
  - $\boxed{E[f(x,y)] = E[(x–\mu_x)(y–\mu_y)] = Cov(x,y)}$, the covariance of x and y
  - Measure how much two variables change together
  - $Cov(x,y)$ positive whenever $x > \mu_x$, then $y > \mu_y$ on average (& vice versa)
  - $Cov(x,y)$ negative whenever $x > \mu_x$, then $y < \mu_y$ on average (& vice versa)

Conditional Expectation

- Remember $p(x|y)$ is conditional probability of x given y?
- Conditional expectation:

$$
\begin{aligned}
E[f(x,y)|y] &\triangleq E_{p(x|y)}[f(x,y)] \\
&\triangleq \sum_x f(x,y)p(x|y) \qquad\qquad \triangleq \int_x f(x,y)p(x|y)dx
\end{aligned}
$$

- Read as "expected value of $f(x,y)$ given y"
- Conditional expectation tells us average value of $f(x,y)$ taking into account $p(x|y)$

N random variables (aka random vector)

- We have focused on 2 random variables x and y
- In real applications, usually more than 2 variables (e.g., photo has > 1M pixels)
- If we observe x_1, x_2, …, x_N multiple times, some combinations of outcomes more likely than others
- This information captured by joint probability distribution function
- Written as $p(x_1, x_2, \ldots, x_N)$, read as probability distribution of x_1 to x_N
- If x_1, x_2, …, x_N are continuous, then p refers to joint probability distribution function (pdf). If discrete, then refers to joint probability mass function (pmf)
- Many properties for two random variables **generalize** naturally to more variables

Lecture 8 – Combining data

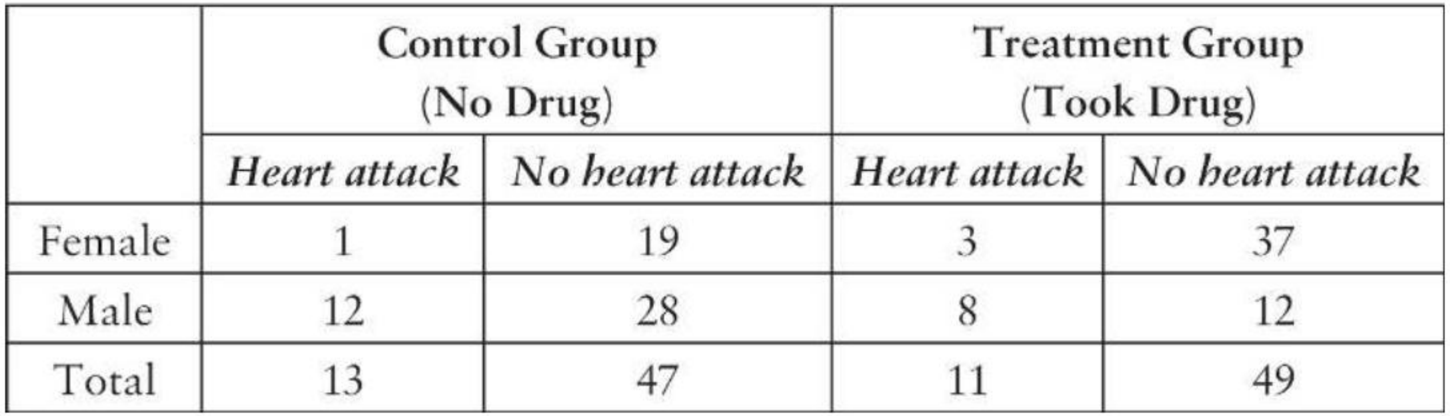
- Combining data
  - Simpson’s paradox in combining data
- Fitting aggregate data
  - Fitting data after Vincent averaging
  - Fitting individual data
- Fitting subgroups of data and individual difference
  - Mixture Modelling
  - K-means Clustering → discrete clusters
  - Structural equation model → continuous along latent space

Combining data

Vote: yes or no

Does the treatment save lives?

Simpson's paradox



	Control Group (No Drug)		Treatment Group (Took Drug)	
	Heart attack	*No heart attack*	*Heart attack*	*No heart attack*
Female	1	19	3	37
Male	12	28	8	12
Total	13	47	11	49

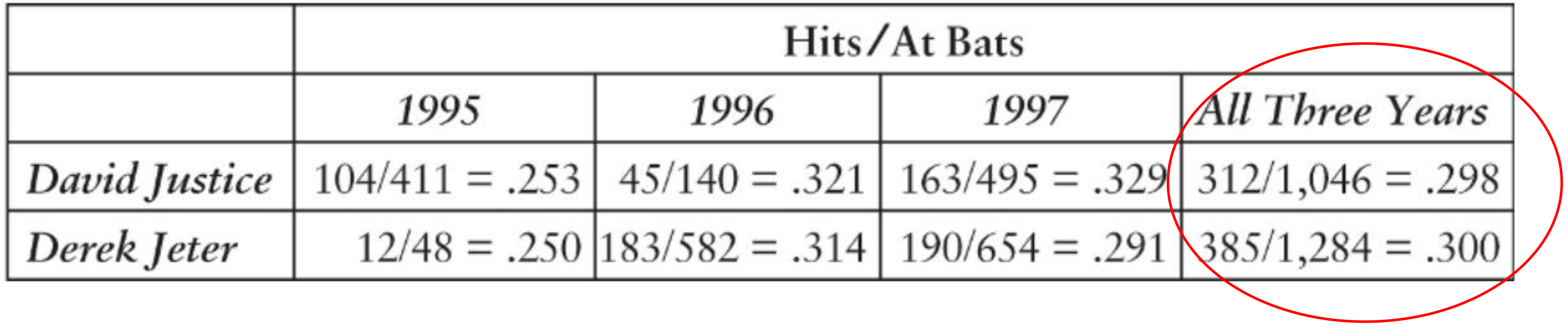
$1/20 < 3/40$

$12/40 < 8/20$

$13/60 > 11/60$

Is *David Justice* better than *Derek Jeter*?

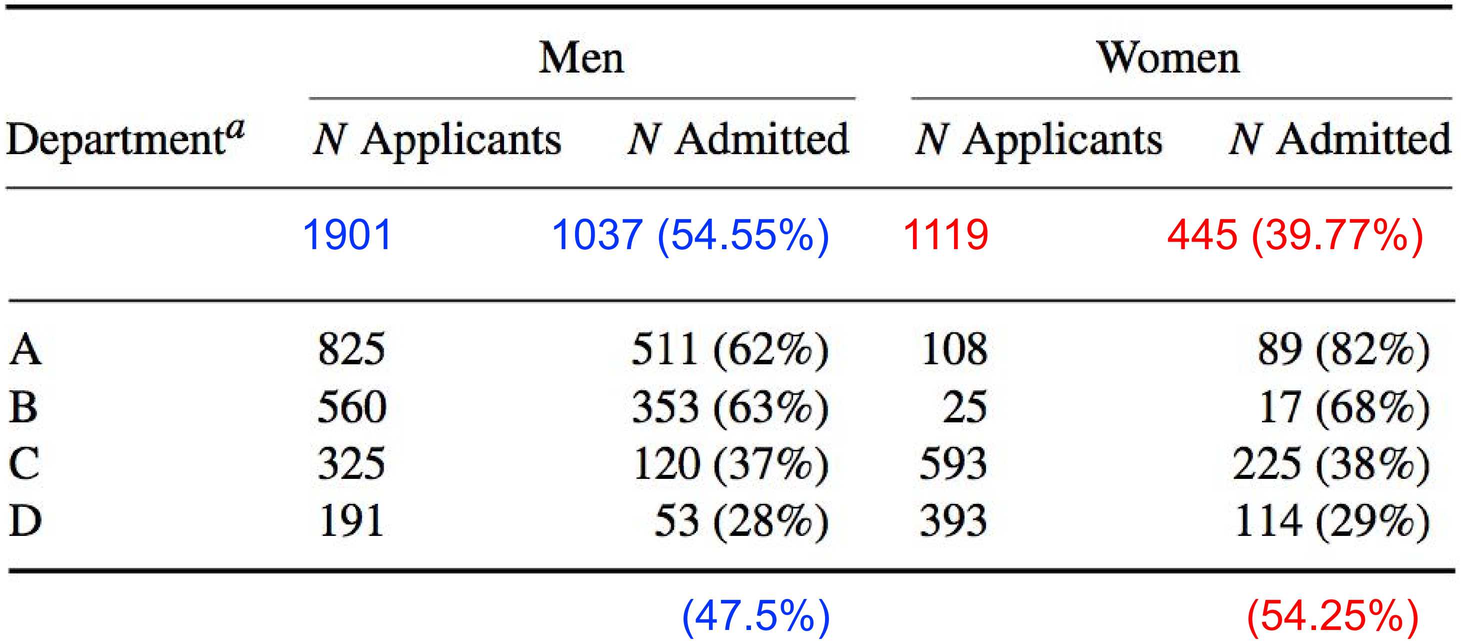
Simpson's reversal



	Hits/At Bats			
	1995	*1996*	*1997*	*All Three Years*
David Justice	104/411 = .253	45/140 = .321	163/495 = .329	312/1,046 = .298
Derek Jeter	12/48 = .250	183/582 = .314	190/654 = .291	385/1,284 = .300

Combining data

Is UC Berkeley gender biased?   Vote: yes or no



Department[a]	Men		Women	
	N Applicants	*N* Admitted	*N* Applicants	*N* Admitted
	1901	1037 (54.55%)	1119	445 (39.77%)
A	825	511 (62%)	108	89 (82%)
B	560	353 (63%)	25	17 (68%)
C	325	120 (37%)	593	225 (38%)
D	191	53 (28%)	393	114 (29%)
		(47.5%)		(54.25%)

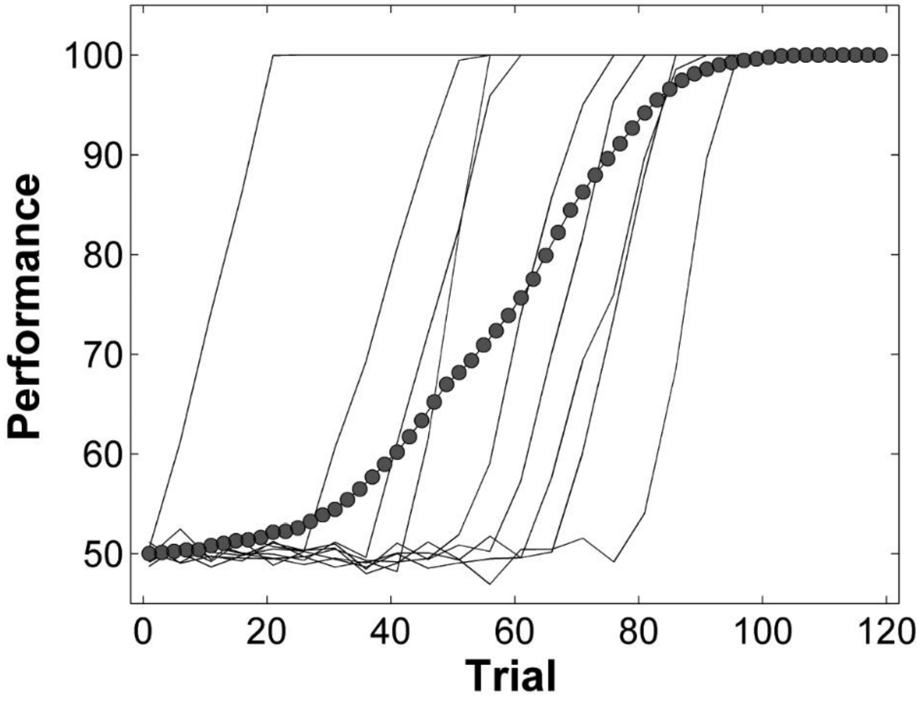
Aggregation of data is **not** always beneficial. It may be misleading.

It matters how you combine data from multiple units!

Implications of Averaging

We usually report results ***at group level***, after averaging the responses from many subjects in a condition.

What could be **wrong** with that?



A learning task

Subjects learn across 120 trials.

For individual data:
Linear learning, starting at chance level (50%) and suddenly starting learning in a linear matter.

For averaged data:
Nonlinear learning, which can fit with a Sigmoid function

When does averaging become a problem?

For a variety of functions, averaging does not present a problem.

- logarithmic functions, $y = a \log x$
- quadratic functions, $y = a + bx + cx^2$

For some functions, shape does change after averaging.

- exponential functions, $y = a + b\, e^{-cx}$

Estes (1956)

Some statistical methods to test **participant heterogeneity.**

When the tests reveal heterogeneity, fitting at the aggregate level is inadvisable.

When the tests fail to detect heterogeneity, fitting at the aggregate level may be permissible.

Smith and Batchelder (2008)

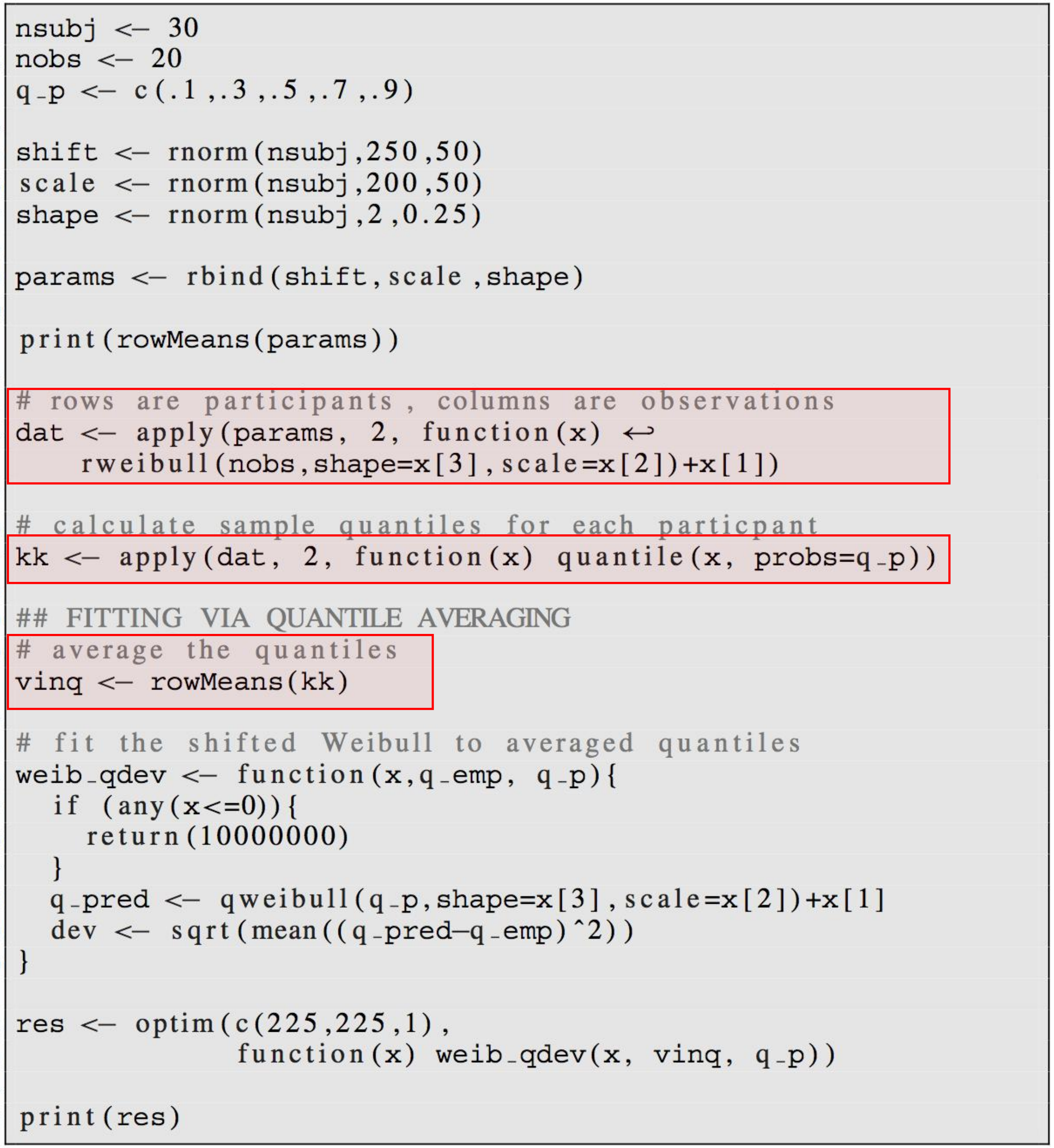
Fitting averaging data

Vincent averaging: an alternative approach to aggregation beyond simple averaging
to retain information about the underlying structure (distribution) of each participant’s responses

Average each quantile across participants.

For example, the 0.1, 0.3, 0.5, 0.7 and 0.9 quantiles correspond to those values that cut off 10%, 30%, 50%, 70%, and 90% of the distribution below (Ratcliff and Smith, 2004)

SEE CODE: lecture8_1_fittingWeibull.R



```r
nsubj <- 30
nobs <- 20
q_p <- c(.1,.3,.5,.7,.9)

shift <- rnorm(nsubj,250,50)
 scale <- rnorm(nsubj,200,50)
shape <- rnorm(nsubj,2,0.25)

params <- rbind(shift,scale,shape)

 print(rowMeans(params))

# rows are participants, columns are observations
dat <- apply(params, 2, function(x) ↩
    rweibull(nobs,shape=x[3],scale=x[2])+x[1])

# calculate sample quantiles for each particpant
kk <- apply(dat, 2, function(x) quantile(x, probs=q_p))

## FITTING VIA QUANTILE AVERAGING
# average the quantiles
vinq <- rowMeans(kk)

# fit the shifted Weibull to averaged quantiles
weib_qdev <- function(x,q_emp, q_p){
  if (any(x<=0)){
    return(10000000)
  }
  q_pred <- qweibull(q_p,shape=x[3],scale=x[2])+x[1]
  dev <- sqrt(mean((q_pred-q_emp)^2))
}

res <- optim(c(225,225,1),
             function(x) weib_qdev(x, vinq, q_p))

 print(res)
```

The Weibull distribution has been used to model response time data in a number of domains.

$$f(x;\lambda,k) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

K: shape parameter
Lambda: scale parameter

Fitting individual data

The **advantages** in fitting individuals' data:

- The sample variability allows us to make inferences about the population.
- If we have a measure of sample variability in each parameter, we can perform *classical tests* (e.g., t-test, ANOVA) to determine whether parameter estimates differ significantly between conditions.

- **SEE CODE**: lecture8_1_fittingWeibull.R

- Disadvantages?

Fitting Subgroups of Data and Individual Differences

For examining heterogeneity in participants:

- **Mixture Modelling** → discrete patterns
- **K-means Clustering** → discrete patterns
- **Structural equation model** → continuous along latent space

Fitting subgroups of data

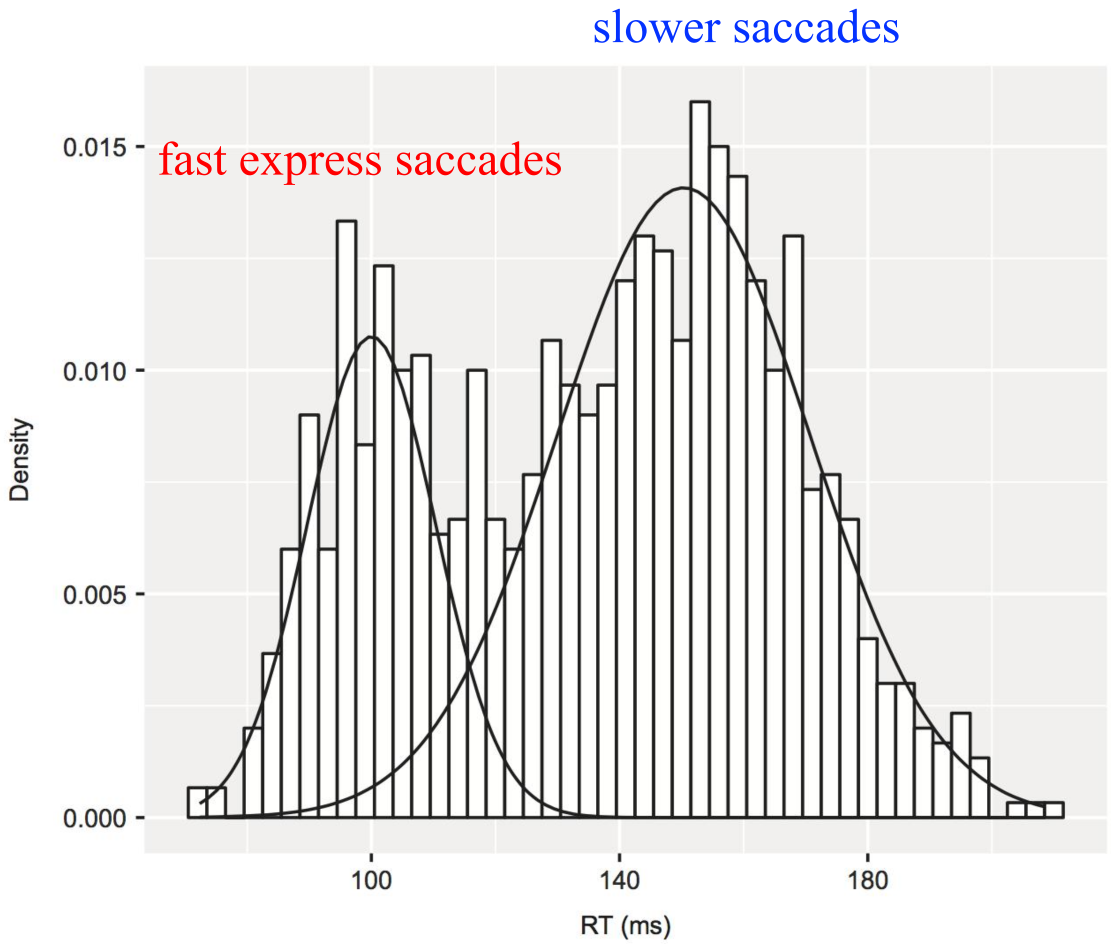
Modelling individual differences

Mixture modeling is useful whenever we expect that our data are obtained from a mixture of different populations or processes.

A mixture model assumes that each data point is sampled from one of N generating models.

Gaussian mixture model

We assume data are sampled from two or more (N) *Gaussian* distributions.



A gap task

- the fixation cross appears shortly before the saccade target appears (the task is to move their eyes to the target).

SEE CODE: (pp113-117)

- lecture8_2_gmmExample.R
- fitting the Gaussian mixture model using the Expectation-Maximization(EM) algorithm

How to determine N?

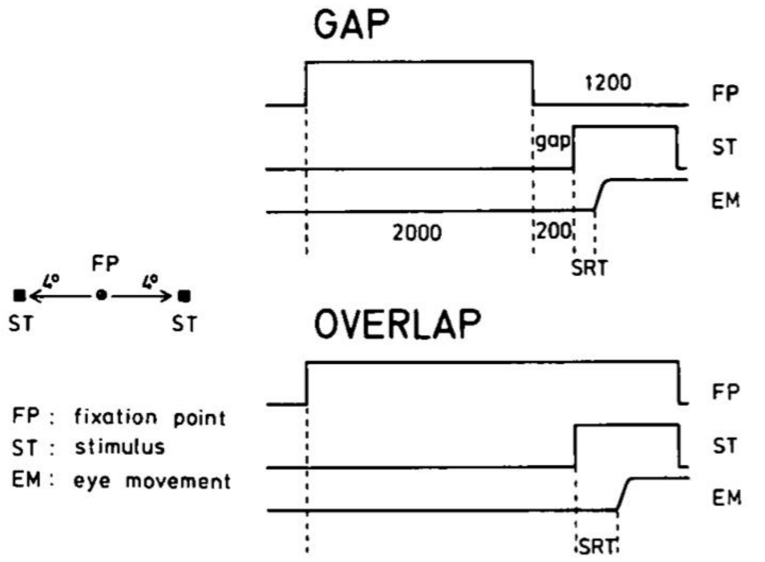
- likelihood ratio tests, AIC, BIC …

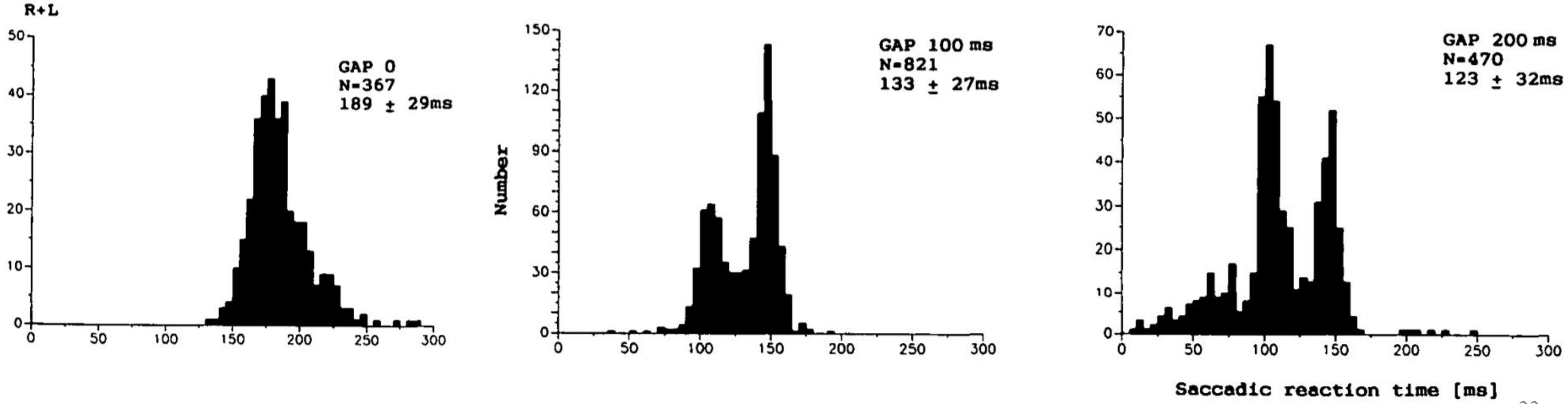
Fischer, B., and Weber, H. (1993)

Eye saccades

A normal adult subject makes 3-5 saccades in a second separated by periods of 200-300 ms during which the eyes do not make large or fast movements.

Hypothesis: the **attentional** system controls vision and eye movements and that it has a dual functional structure which is under different amounts of voluntary control depending on the amount of practice the subjects have and on the state of maturation of the brain.





Fischer, B., and Weber, H. (1993)

Saccade size

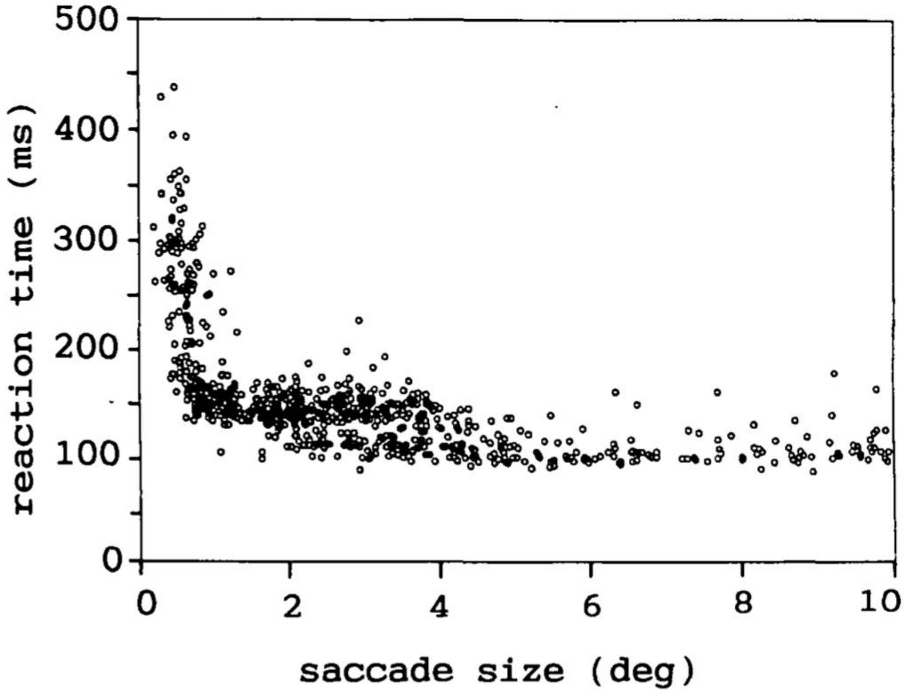


Figure 8. Scatter plot of saccade size versus reaction time. Express saccades can be seen between 2 and 10 deg. Fast regular saccades occur between 0.5 and 4 deg. Below 0.5 deg reaction times increase drastically. The data are taken from Weber et al. (1992).

Age effects

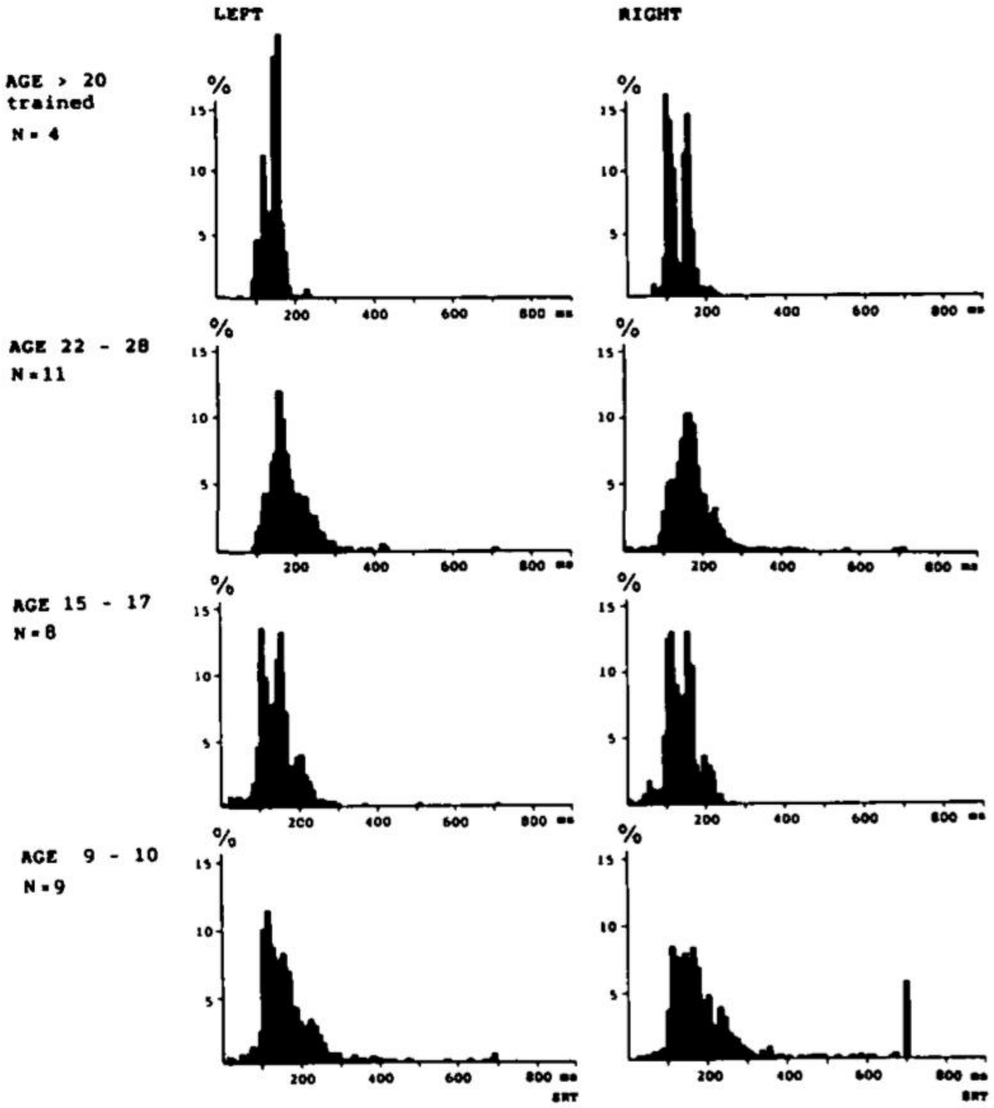


Figure 10. Effect of age on the distribution of saccadic reaction time. Note the strong asymmetry between right- and left-directed saccades for the adult group (upper panel). In all cases the target was randomly presented to the right or left.

Dyslexic readers

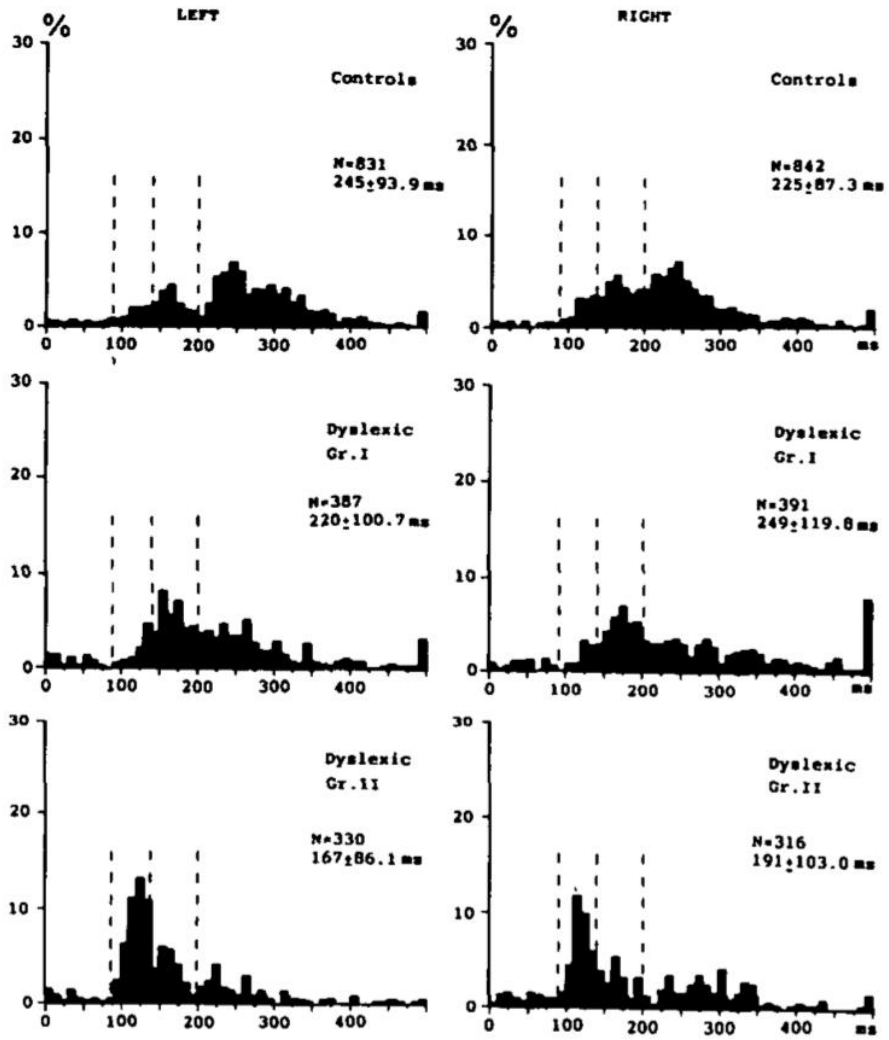
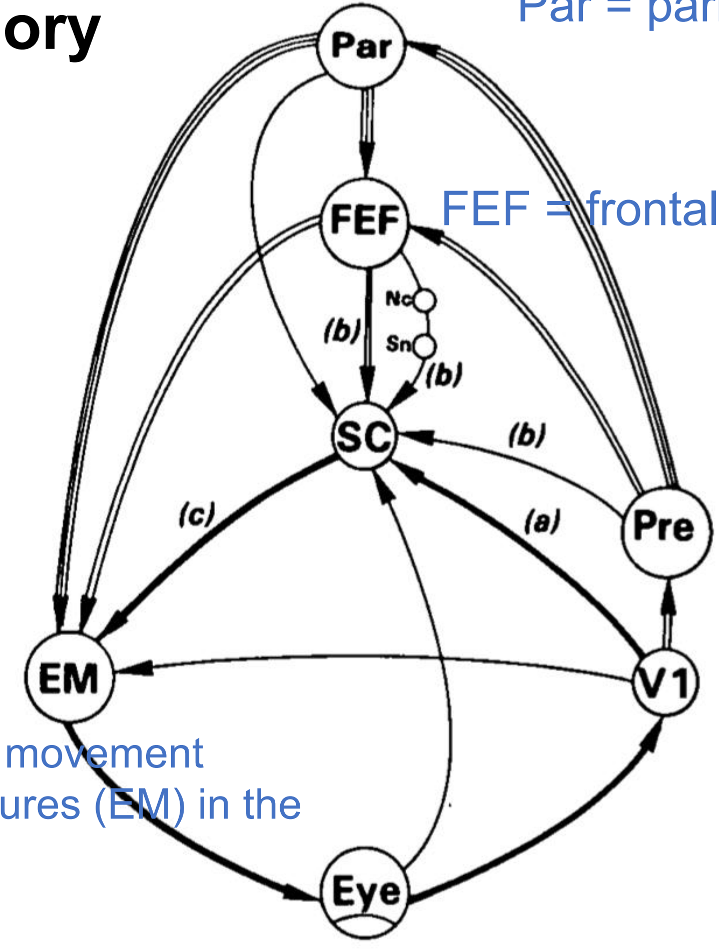


Figure 11. Saccadic reaction times of normal readers and two groups of dyslexic readers obtained in overlap trials. Note the big population of express saccades in the group II dyslexics. The saccade target was presented randomly at 4 deg to the right or left of the fixation point.

Theory



Par = parietal cortex

FEF = frontal eye field

EM, efferent eye movement generating structures (EM) in the brain stem

VI = striate visual cortex (the lateral geniculate nucleus is omitted), Pre = prestriate visual cortex including areas V2 and V5 and area MST,

SC = superior colliculus,

Nc = nucleus caudatus,

Sn = substantia nigra pars reticulata.

K-Means Clustering

K-means clustering is an unsupervised learning. It assumes that data belong to one of several discrete clusters.

The clusters are defined by *centroids* (i.e., cluster centers).

The aim of the K-means algorithm is to assign individual objects to the clusters so as to minimize the within-cluster sum of squares (i.e., the sum of squares between objects and the centroid of the cluster to which they are assigned).

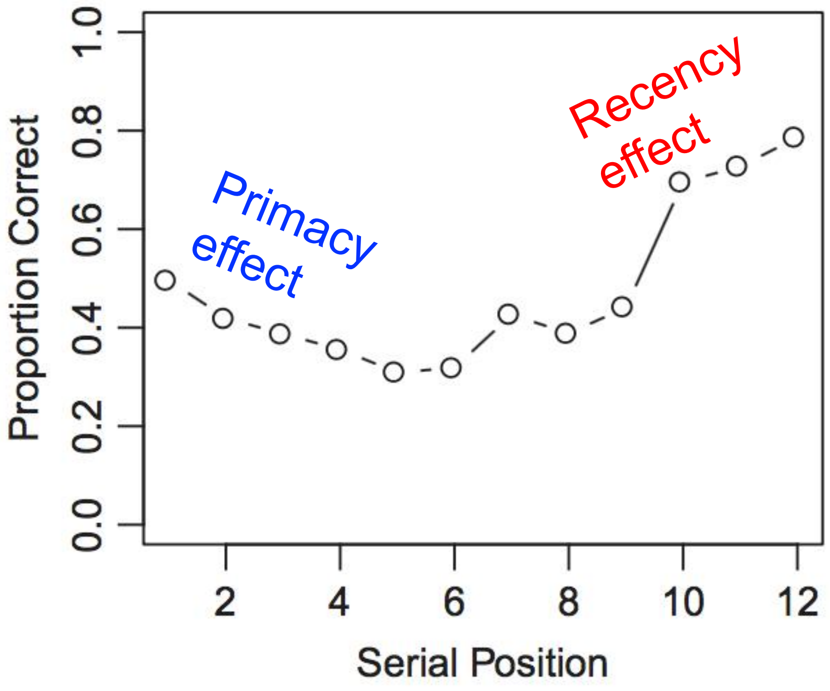
1. Specify the number of clusters (*K*). For each cluster, specify an initial centroid (a vector).
2. **Assign** each object to the **cluster** to which it is closest. The distance between each object and each centroid is usually measured using Euclidean distance.
3. **Recalculate** each cluster centroid by averaging across all objects that have been assigned to that cluster.
4. Keep **repeating *steps 2&3*** until the assignment of objects to clusters no longer changes.

K-Means Clustering

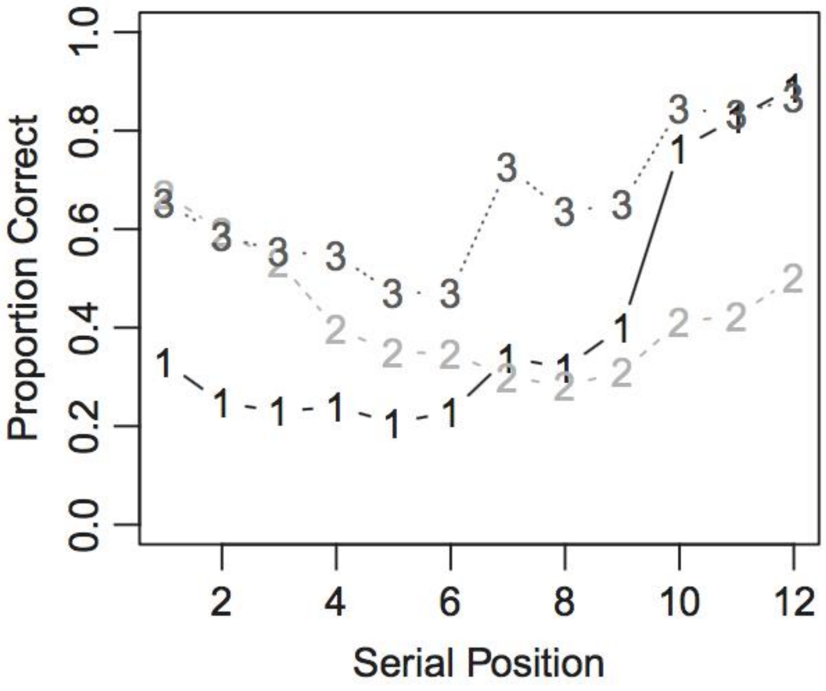
An episodic memory task:

Present 12 words, and then recall those words in a specific order.

Group level



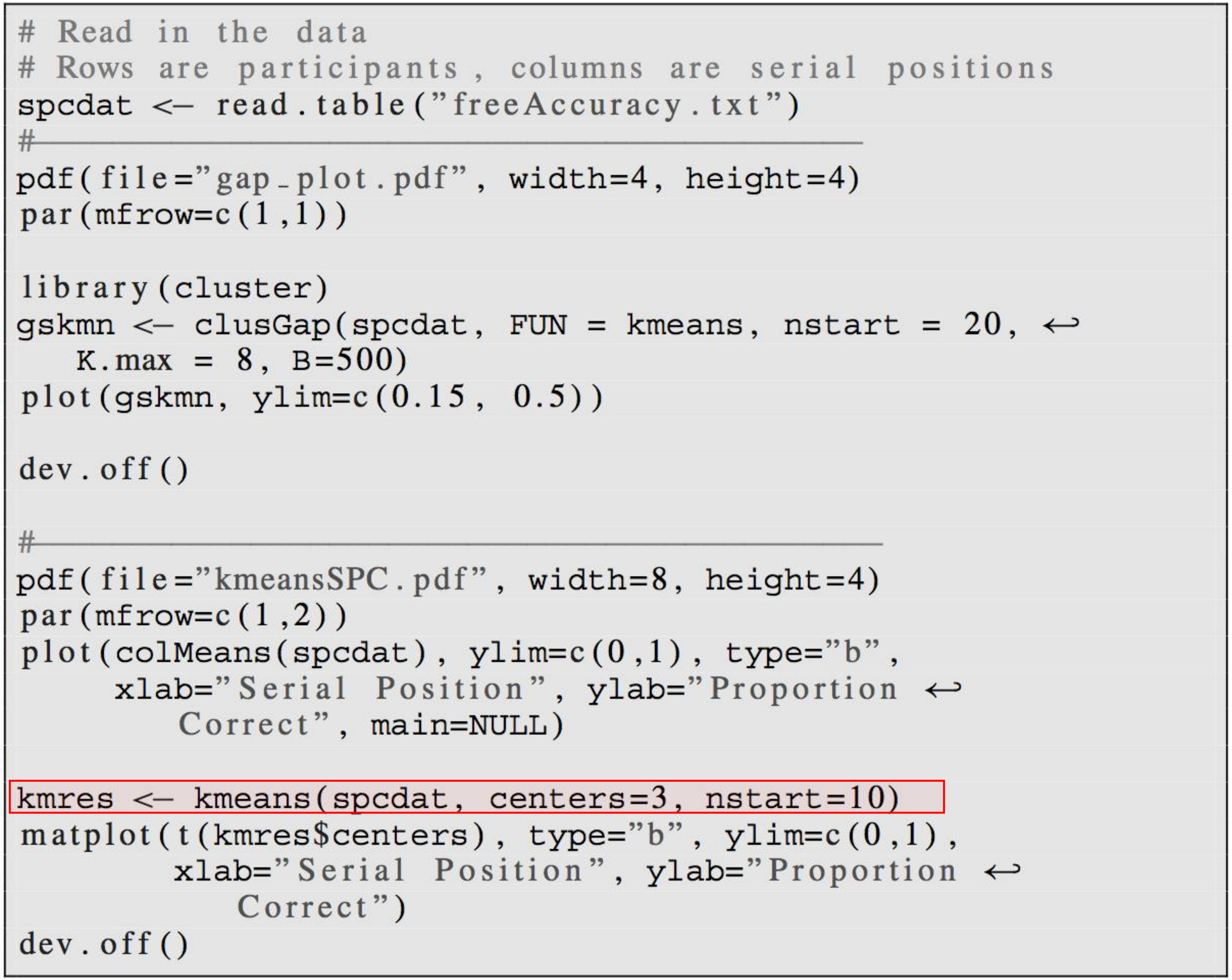
3 clustering patterns



Pattern 1: major recency

Pattern 2: major primacy

Pattern 3: both



```
# Read in the data
# Rows are participants, columns are serial positions
spcdat <- read.table(”freeAccuracy.txt”)
#———————————————————————————————
pdf(file=”gap_plot.pdf”, width=4, height=4)
par(mfrow=c(1,1))

library(cluster)
gskmn <- clusGap(spcdat, FUN = kmeans, nstart = 20, ↩
    K.max = 8, B=500)
plot(gskmn, ylim=c(0.15, 0.5))

dev.off()

#———————————————————————————————
pdf(file=”kmeansSPC.pdf”, width=8, height=4)
par(mfrow=c(1,2))
plot(colMeans(spcdat), ylim=c(0,1), type=”b”,
     xlab=”Serial Position”, ylab=”Proportion ↩
         Correct”, main=NULL)

kmres <- kmeans(spcdat, centers=3, nstart=10)
matplot(t(kmres$centers), type=”b”, ylim=c(0,1),
         xlab=”Serial Position”, ylab=”Proportion ↩
             Correct”)
dev.off()
```

How to determine K?

The gap statistic was introduced by Tibshirani et al. (2001) as a method of determining an appropriate number of clusters to characterize a data set.

The algorithm works by determining, for each value of *k*, the difference (gap) between the observed **within-cluster sum of squares**, and that expected under some **null reference model**.

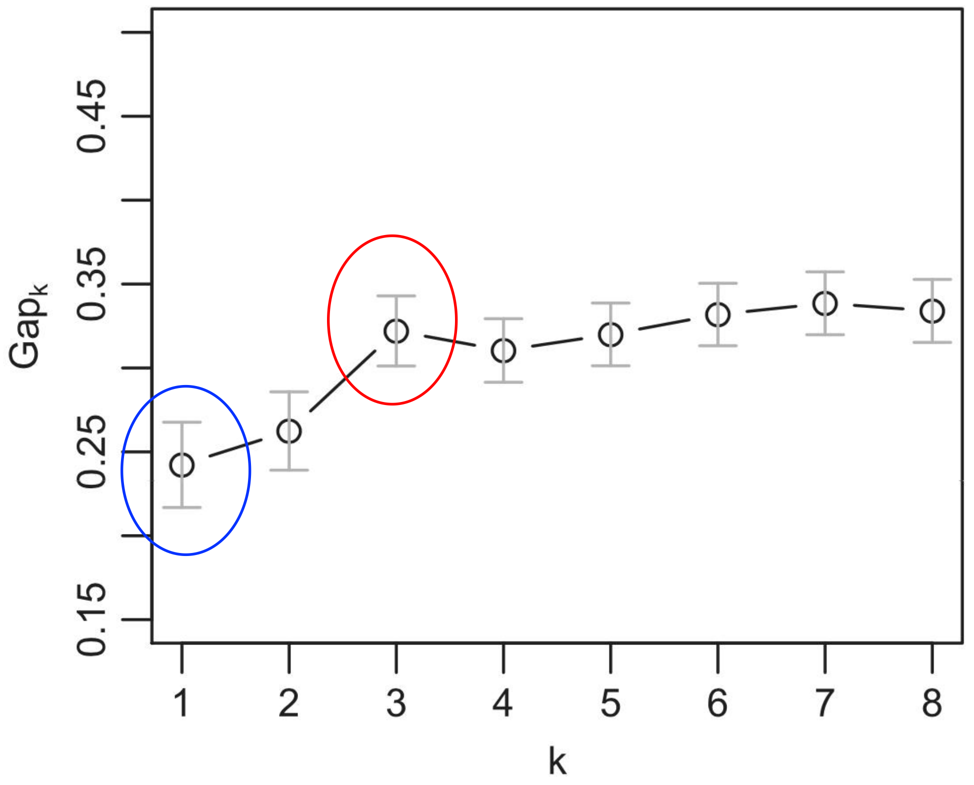
The expectation under the null model is determined by bootstrapping from the null model.

In the simplest version, we sample each feature uniformly from the range of values observed for that feature in the data set.

$$Gap(k) \geq Gap(k+1) - s_{k+1}$$

```
library(cluster)
gskmn <- clusGap(spcdat, FUN = kmeans, nstart = 20,
K.max = 8, B=500)
plot(gskmn, ylim=c(0.15, 0.5))
```

The gap statistic for different value k



Tibshirani et al. (2001)

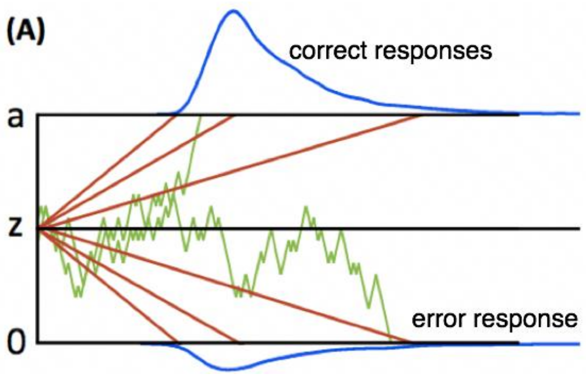
Modeling Individual Differences

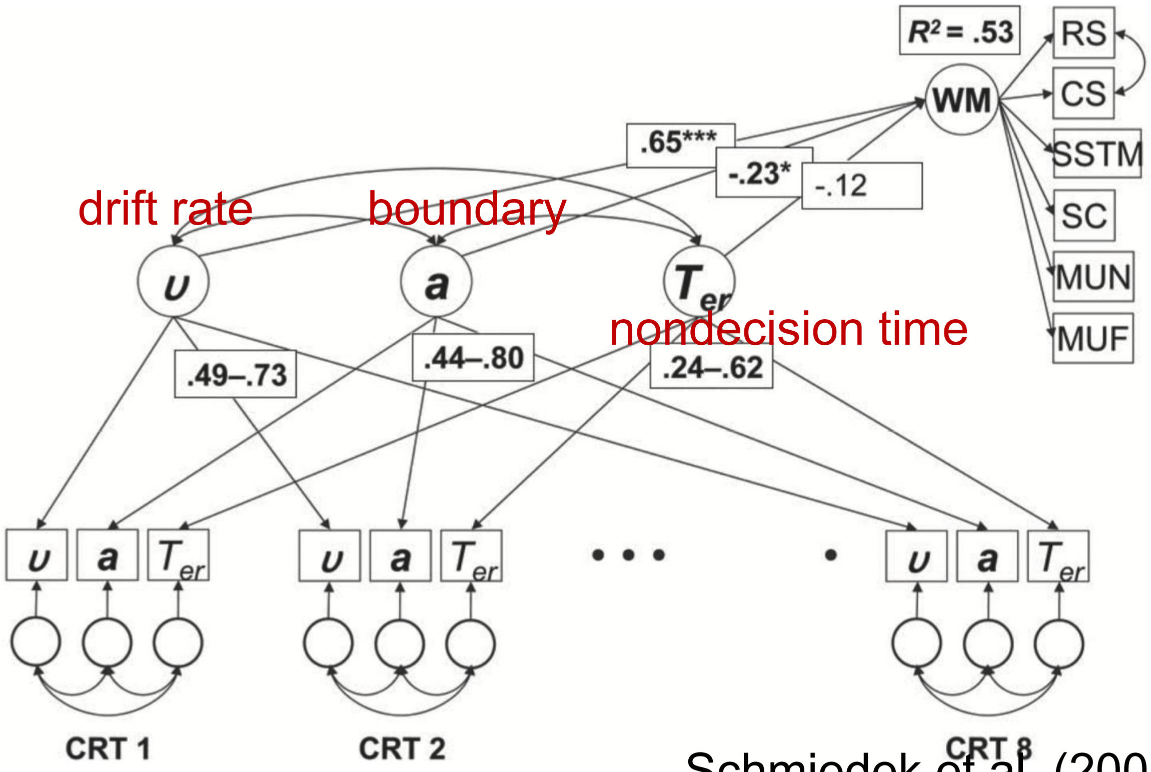
- Differences between individuals are not always manifested as discrete clusters.
- Textbook, PP121-122
- **Structural equation model (SEM)** → continuous along latent space

SEM concurrently estimates:

a) the correlations between latent and observed (manifest) variables,

b) the correlations between the latent variables.





drift rate

boundary

nondecision time

Schmiedek et al. (2007)

Lecture 8 – Summary

- Combining data
  - Simpson's paradox in combining data
  - Simply averaging or aggregating may be problematic!
- Fitting aggregate data
  - Fitting data after Vincent averaging
  - Fitting individual data
- Fitting subgroups of data and individual difference
  - Mixture Modelling
  - K-means Clustering → discrete clusters
  - Structural equation model → continuous along latent space

Reading materials

Textbooks (must read)

- **Chapter 5 (Combining Information from Multiple Participants)**

Extra reading (for fun)

- The book of why by Judea Pearl– chapter 6 (lots of examples of paradoxes)

Papers (details about the examples in the course)

- Fischer, B., and Weber, H. (1993). Express saccades and visual attention. Behavioral and Brain Sciences, 16, 553–567.
- Schmiedek, F., et al. (2007). Individual differences in components of reaction time distributions and their relations to working memory and intelligence. Journal of Experimental Psychology: General, 136, 414–429.