Chapter 10 Design

10.1 Overview of experimental design

According to R. A. Fisher (The Design of Experiments), the basic principles of experimental design are

- Replication: deals with variation/uncertainty, allows for generalization
- Randomization: deals with confounding factors, allows for cause-effect statements
- Blocking: reduces known but irrelavant sources of variation, improves statistical efficiency

An experiment in the real world may have many more complications. We consider a simple case where we want to consider the comparison of a numerical response under different scnearios. Suppose that we want to study whether treatment A causes any changes in the reponse. Consider the following scenarios

1. We conduct a survey to measure the response and Treatment A in the target population.
2. We enroll participants from the target population, and randomly assign Treatment A and placebo to the participants.

Scenario 1 is an **observational study**, given that the data is "observed". Scenario 2 is a **controlled randomized experiment (or trial)**. In general, we can only infer the association between the treatment and response from an observational study, but a controlled randomized experiment can lead to causal inference. See Chapter 9 for a more rigorous discussion.

However, regardless of the design, the statistical analysis strategy is almost the same in both scenarios. We group observations into two groups based on their treatment status, the treatment group and the control group. The inference of association or cause-effect can be done by a statistical comparison between the two group.

Usually the goal of a study is to find out which and how explanatory factors influence/relate to the response variables.

The design of a study consists of making decisions on the following:

- The set of explanatory factors
- The set of response variables
- The set of treatments
- The set of experimental units
- The way of randomization and blocking
- Sample size and number of replications
- ...

The **factors** are the explanatory variables that must contain the **variables of interest**.

Examples:

- In a study of the effects of colors and prices on sales of cars, the factors being studied are color (qualitative variable) and price (quantitative variable)

- In an investigation of the effects of education on income, the factor being studied is education level (What type is it?).

The **factor levels** are the values of the factor.

Examples: Sex has two levels: male, female

There are a various types of factors. One way of classification is to classify them based on the experimental design.

- Experimental factors: Levels of the factor are assigned at random to the experimental units

- Observational factors: Levels of the factor are characteristic of the experimental units and is not under the control of the investigators. Note that there could be observational factors in an experimental study.

Examples: Consider an RCT on the effect of a new drug. If we are also interested in the effects of age and gender on the recovery rate, then these are observational factors; While the new drug is an experimental factor.

In a single factor study, each treatment corresponds to a factor level The number of treatments equals to the number of levels of that factor In a multi-factor study, each treatment corresponds to a combination of factor levels across different factors. The number of possible treatments is the multiplication of the number of factor levels across different factors.

Examples:

- In the study of effects of education on income, each education level is a treatment (high school, college, advanced degree, etc.)
- In the study of effects of race and gender on income, each combination of race and gender is a treatment (Asian female; Hispanic male, etc.)

Q: How many possible treatments are there for a study with two factors, one with three levels and another with four levels?

Levels of each factor is another important aspect to consider when designing a study. For qualitative factors, levels are indicated by the nature of the factor (e.g., sex has two levels: female and male). For quantitative factors, the choice of levels reflects the type of trend expected by the investigator (e.g., linear trend <--> two levels; quadratic trend <--> three levels). Usually 3 to 4 equally spaced levels are sufficient. The range of levels are also crucial. prior knowledge is often required for an effective choice of factors and treatments.

An **experimental unit** is the smallest unit of experimental material to which a treatment can be assigned. The definition of an experimental unit is determined by the way of randomization.

Example: In a study of two retirement systems involving the 10 UC schools, we would ask if the basic unit should be an individual employee, a department, or a University. The basic unit should be an entire University for practical feasibility.

Example: A study conducted surveys among 5,000 US college students, and found out that about 20% of them had used marijuana at least once. If the goal is to study drug usage among

Americans aging from 18 to 22, is this a good design? The experimental units should be representative of the population about which conclusions are going to be made.

Sample size is the **total number** of experimental units in the study. Sample size is usually determined by the trade-off between (i) statistical considerations such as power of tests, precision of estimations and (ii) the availability of resources such as money, time, man power. In general, the larger the sample size, the better for statistical inference; However, the more costly the study is

For many designed studies, the sample size is an integer multiple of the number of treatments. This integer is the number of times each treatment being repeated. One complete repetition of all treatments (under similar experimental conditions) is called a complete replicate of the experiment.

When a treatment is repeated under the same experimental conditions, any differences in the responses are due to random fluacuations. Thus replication provides us information about the degree of random fluctuation. If the variation due to random fluctuation is relatively small compared to the total variation in the responses, we would have evidence for treatment effect.

10.2 Randomized experiments

10.2.1 Simple randomized experiments

Suppose that we want to study whether treatment A causes any changes in the reponse. We know from the potential outcome framework that independence between the treatment assignment and the potential outcomes is key to average treatment/causal effect.

Even for a simple randomized experiment, there are different ways to assign the treatment.

1. For any subject, $i = 1, \dots, n$, randomize its treatment $Z_i = 0$ or 1 with a probability p.
2. For $i = 1, \dots, n$, randomly allocate them into two groups with $n/2$ in each group.

Unit-level treatment effect as $\tau_i(j, j') = Y_i(j) - Y_i(j')$.

Population-level treatment effect is

$$\tau(j, j') = N^{-1} \sum_{i=1}^{n} \tau_i(j, j') = N^{-1} \sum_{i=1}^{N} \{Y_i(j) - Y_i(j')\} \equiv \bar{Y}(j) - \bar{Y}(j').$$

Fisher's sharp null hypothesis of zero individual treatment effects

$$H_{0F} : Y_i(1) = Y_i(2) = \cdots = Y_i(J) \;\; (i = 1, \dots, N)$$

This leads to the permutation test.

Neyman's null hypothesis: no average treatment effects.

$$H_{0N} : \bar{Y}(1) = \cdots = \bar{Y}(J).$$

Weaker restriction on the potential outcomes.

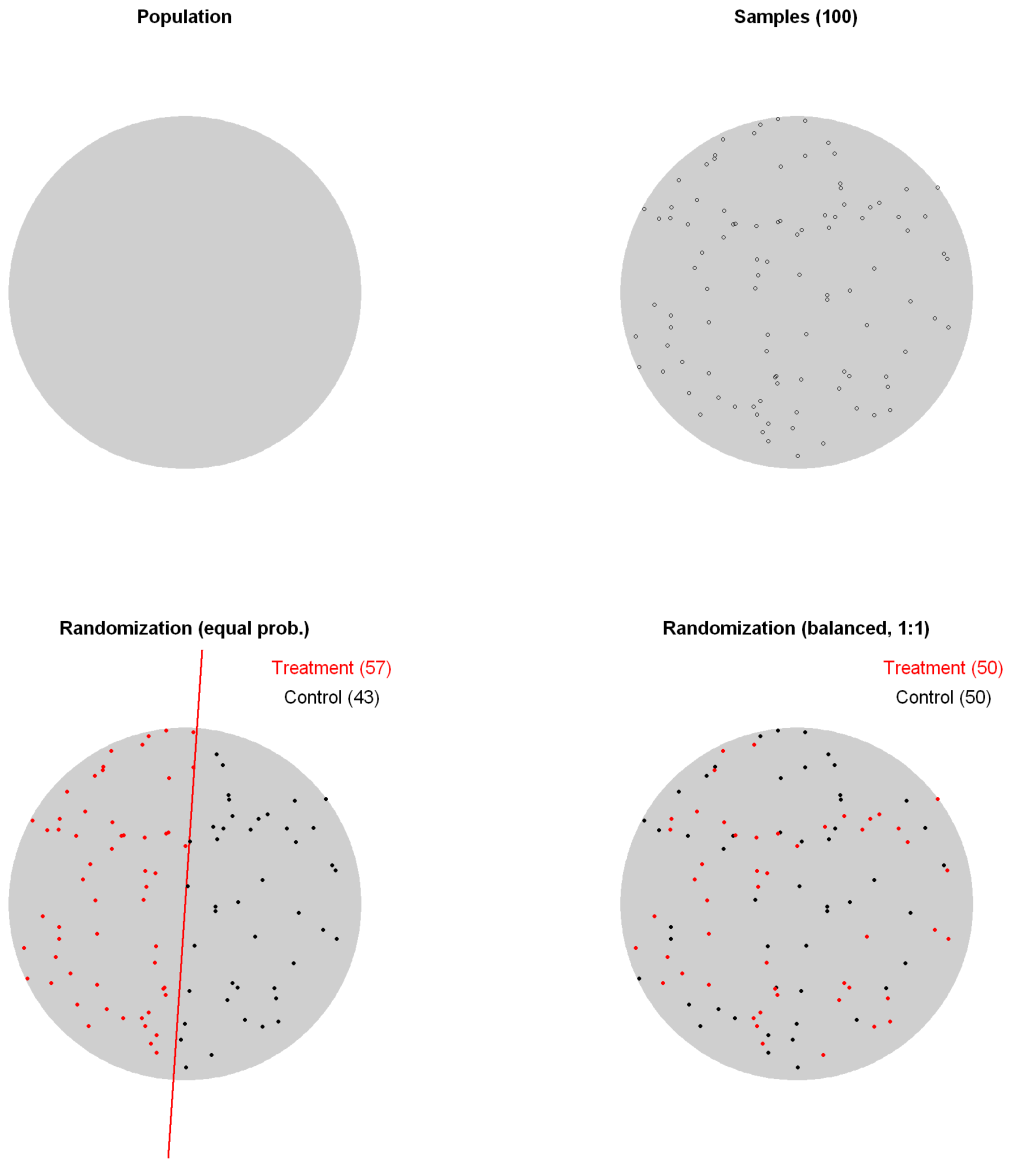
Claim: Fisher's randomization test fails to control type I error in unbalanced experiments.

Example: Visualization of simple randomization

In [1]:

```
center.coord=c(45,45)

cols=c('black','red')
R=3

n=100

options(repr.plot.width=18, repr.plot.height=18)
par(mfrow=c(2,2))
plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',yaxt='n',xlab='',ylab='',bty='n',
    main='Population',cex.main=2)
theta <- seq(0,2*pi, length=400)
x1 <- R*cos(theta)+center.coord[1]
y1 <- R*sin(theta)+center.coord[2]

polygon(x1,y1, border=NA,col="#00000030")


plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',yaxt='n',xlab='',ylab='',bty='n',
    main=paste0('Samples (', n, ')'),cex.main=2)
theta <- seq(0,2*pi, length=400)
x1 <- R*cos(theta)+center.coord[1]
y1 <- R*sin(theta)+center.coord[2]

polygon(x1,y1, border=NA,col="#00000030")

# Add dots on the plot
r = R*sqrt(runif(n))
theta = runif(n) * 2 * pi
x = center.coord[1] + r * cos(theta)
y = center.coord[2] + r * sin(theta)
points(x=x,y=y,pch=1,col='black')


plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',yaxt='n',xlab='',ylab='',bty='n',
    main='Randomization (equal prob.)',cex.main=2)
theta <- seq(0,2*pi, length=400)
x1 <- R*cos(theta)+center.coord[1]
y1 <- R*sin(theta)+center.coord[2]

polygon(x1,y1, border=NA,col="#00000030")

# Treatment and control:
slope=20*(runif(1)-0.5)+10
abline(a=center.coord[2]-slope* center.coord[1],b=slope,col=cols[2],lwd=3  )
trt=1+((y - center.coord[2]) > slope*(x-center.coord[1]))
points(x=x,y=y,pch=16,col=cols[trt])

text(x=center.coord[1]+R-0.5,y=center.coord[2]+R+1,labels=paste0('Treatment (', sum(t
text(x=center.coord[1]+R-0.5,y=center.coord[2]+R+0.5,labels=paste0('Control (', sum(t

plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',yaxt='n',xlab='',ylab='',bty='n',
    main='Randomization (balanced, 1:1)',cex.main=2)
theta <- seq(0,2*pi, length=400)
x1 <- R*cos(theta)+center.coord[1]
y1 <- R*sin(theta)+center.coord[2]
```

```
polygon(x1,y1, border=NA,col="#00000030")

# Treatment and control:
trt=rep(1,n)
trt[sample(1:n, n/2)]=2
points(x=x, y=y,pch=16,col=cols[trt])

text(x=center.coord[1]+R-0.5,y=center.coord[2]+R+1,labels=paste0('Treatment (', sum(t
text(x=center.coord[1]+R-0.5,y=center.coord[2]+R+0.5,labels=paste0('Control (', sum(t
par(mfrow=c(1,1))
```



10.2.2 Stratified randomized experiment

This is also known as randomized block design. The blocking can improve efficiency, and sometimes is required for feasibility. Stratification during design can be seen as covariate adjustment in a randomized experiment.

Recall that $\text{ACE} \equiv \mathbb{E}[Y(1) - Y(0)]$ and $\hat{\text{ACE}} = \bar{Y}_1 - \bar{Y}_0$.

Let X_i be the centered covariates and $\mathbb{E}[X_i] = 0$, $\bar{X}_z = N_z^{-1} \sum_{i=1}^{N} X_i 1[Z_i = z]$, and $\mathbb{E}[\bar{X}_z | Z] = 0$. Finally, we assume that $Z_i \perp X_i$.

Consider a new estimator

$$\hat{\mathrm{ACE}}(\gamma_1,\gamma_0) = \left(\bar{Y}_1 - \gamma_1^T\bar{X}_1\right) - \left(\bar{Y}_0 - \gamma_0^T\bar{X}_0\right).$$

We can verify that $\mathbb{E}\left[\hat{\mathrm{ACE}}(\gamma_1,\gamma_0)\right] = \mathrm{ACE}$ for any γ_1 and γ_0. Therefore, we have seen that the new estimator is an unbiased estimator, now our task is to find the best γ_0 and γ_1 to minimize its variance. We can use the linear projection idea in Chapter 9.

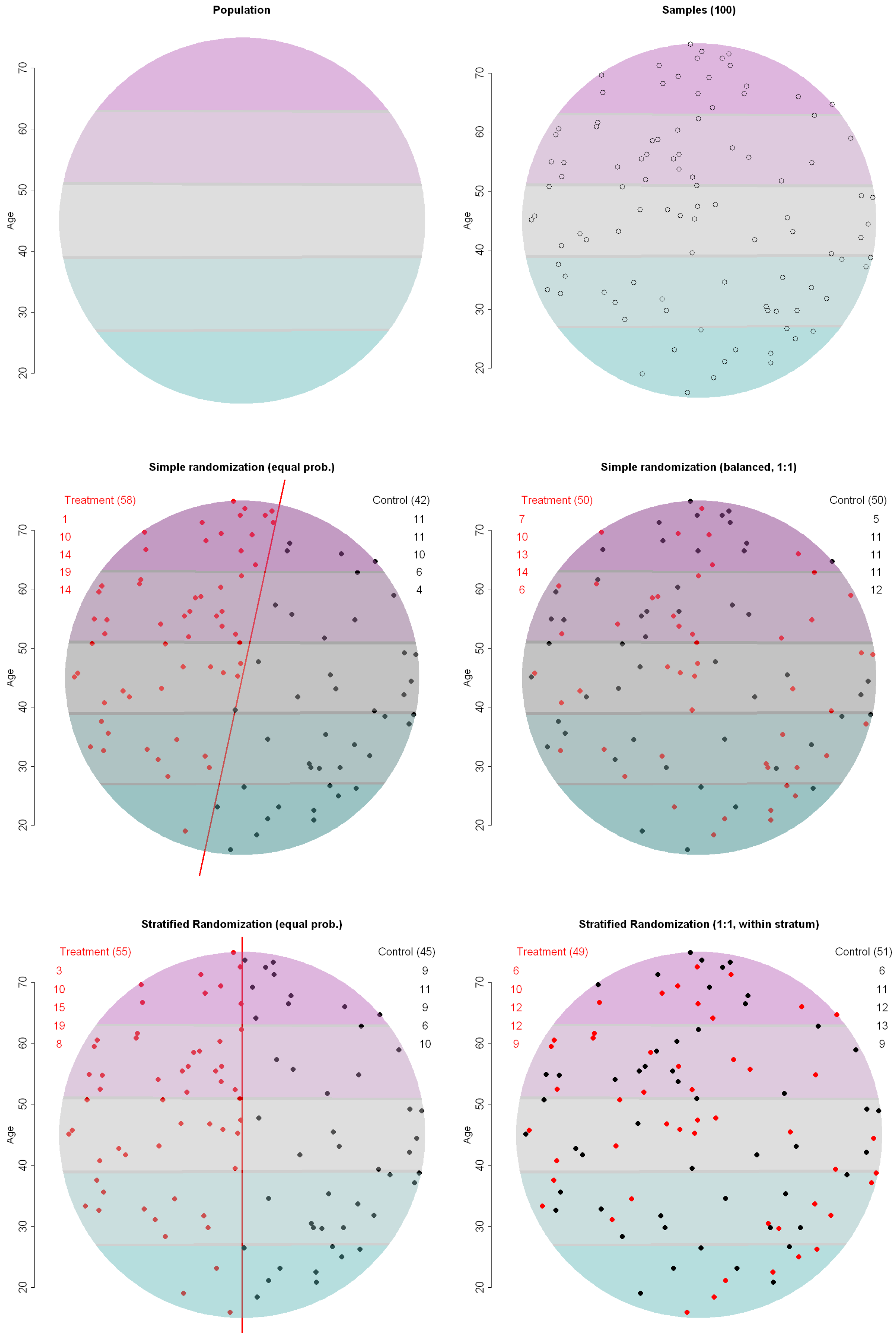
Example: Visualization of stratification

In [2]:

```
center.coord=c(45,45)

cols=c('black','red')
R=30
n=100

# Strata:

n.strata=5
strata.bounds=seq(from=center.coord[2]-R,to=center.coord[2]+R,length.out=n.strata+1)
strata.cols=paste0(cm.colors(n.strata),50)


options(repr.plot.width=18, repr.plot.height=27)
par(mfrow=c(3,2))
plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*R,ylim=center.coord[2]+c(-1,1)*R,
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main='Population',cex.main=2)
theta <- seq(0,2*pi, length=400)
x1 <- R*cos(theta)+center.coord[1]
y1 <- R*sin(theta)+center.coord[2]

polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which( y1>=lwb & y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')
}

plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main=paste0('Samples (', n, ')'),cex.main=2)

polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which( y1>=lwb & y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')
}

# Add dots on the plot
r = R*sqrt(runif(n))
theta = runif(n) * 2 * pi
x = center.coord[1] + r * cos(theta)
y = center.coord[2] + r * sin(theta)
points(x=x,y=y,pch=1,col='black',cex=2)



# Simple:
plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main='Simple randomization (equal prob.)',cex.main=2)
```

```
polygon(x1,y1, border=NA,col="#00000030")

# Treatment and control:
slope=20*(runif(1)-0.5)+10
abline(a=center.coord[2]-slope* center.coord[1],b=slope,col=cols[2],lwd=3  )
trt=1+((y - center.coord[2]) > slope*(x-center.coord[1]))
points(x=x,y=y,pch=16,col=cols[trt],cex=2)

strata.counts=matrix(0,nrow=n.strata,ncol=2)
polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which(  y1>=lwb &  y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')

    tmp=trt[ y>lwb &  y<=upb ]
    strata.counts[i,]=c(sum(tmp==1),sum(tmp==2))
}


text(x=center.coord[1]-R*(0.8),y=center.coord[2]+R,labels=paste0('Treatment (', sum(t

text(x=center.coord[1]+R*(0.9),y=center.coord[2]+R,labels=paste0('Control (', sum(trt

for(i in 1:n.strata){
  text(x=center.coord[1]-R,y=center.coord[2]+R*(1-i*0.1),labels=strata.counts[i,2],co
  text(x=center.coord[1]+R,y=center.coord[2]+R*(1-i*0.1), labels=strata.counts[i,1],c
  }



plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*(R+1),ylim=center.coord[2]+c(-1,1)*(
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main='Simple randomization (balanced, 1:1)',cex.main=2)

polygon(x1,y1, border=NA,col="#00000030")

# Treatment and control:
trt=c(rep(1,n/2),rep(2,n/2))
points(x=x,y=y,pch=16,col=cols[trt],cex=2)

strata.counts=matrix(0,nrow=n.strata,ncol=2)
polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which(  y1>=lwb &  y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')

    tmp=trt[ y>lwb &  y<=upb ]
    strata.counts[i,]=c(sum(tmp==1),sum(tmp==2))
}


text(x=center.coord[1]-R*(0.8),y=center.coord[2]+R,labels=paste0('Treatment (', sum(t

text(x=center.coord[1]+R*(0.9),y=center.coord[2]+R,labels=paste0('Control (', sum(trt

for(i in 1:n.strata){
  text(x=center.coord[1]-R,y=center.coord[2]+R*(1-i*0.1),labels=strata.counts[i,2],co
  text(x=center.coord[1]+R,y=center.coord[2]+R*(1-i*0.1), labels=strata.counts[i,1],c
  }
```

```
# Stratification:

plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*R,ylim=center.coord[2]+c(-1,1)*R,
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main='Stratified Randomization (equal prob.)',cex.main=2)

# Treatment and control:
abline(v=center.coord[2],col=cols[2],lwd=3  )
trt=1+( (x-center.coord[1])<0)
points(x=x,y=y,pch=16,col=cols[trt],cex=2)

strata.counts=matrix(0,nrow=n.strata,ncol=2)
polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which(  y1>=lwb &  y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')

    tmp=trt[ y>lwb &  y<=upb ]
    strata.counts[i,]=c(sum(tmp==1),sum(tmp==2))
}


text(x=center.coord[1]-R*(0.8),y=center.coord[2]+R,labels=paste0('Treatment (', sum(t

text(x=center.coord[1]+R*(0.9),y=center.coord[2]+R,labels=paste0('Control (', sum(trt

for(i in 1:n.strata){
  text(x=center.coord[1]-R,y=center.coord[2]+R*(1-i*0.1),labels=strata.counts[i,2],co
  text(x=center.coord[1]+R,y=center.coord[2]+R*(1-i*0.1), labels=strata.counts[i,1],c
  }


plot(0, 0, type='n',xlim=center.coord[1]+c(-1,1)*R,ylim=center.coord[2]+c(-1,1)*R,
     asp=1,xaxt='n',xlab='',ylab='Age',bty='n',cex.axis=2,cex.lab=1.8,
    main='Stratified Randomization (1:1, within stratum)',cex.main=2)

# Treatment and control:

trt=rep(1,n)
strata.counts=matrix(0,nrow=n.strata,ncol=2)
polygon(x1,y1, border=NA,col="#00000030")
for(i in 1:n.strata){
    lwb=strata.bounds[i];upb=strata.bounds[i+1]
    idx=which(  y1>=lwb &  y1<=upb);
    polygon(x1[idx],y1[idx], col=strata.cols[i],border='NA')

    idx=which(y>lwb &  y<=upb )
    n.s=length(idx)
    trt[ sample(idx,round(n.s/2)) ]=2
    strata.counts[i,]=c(sum(trt[idx]==1),sum(trt[idx]==2))
}

points(x=x,y=y,pch=16,col=cols[trt],cex=2)

text(x=center.coord[1]-R*(0.8),y=center.coord[2]+R,labels=paste0('Treatment (', sum(t

text(x=center.coord[1]+R*(0.9),y=center.coord[2]+R,labels=paste0('Control (', sum(trt

for(i in 1:n.strata){
  text(x=center.coord[1]-R,y=center.coord[2]+R*(1-i*0.1),labels=strata.counts[i,2],co
  text(x=center.coord[1]+R,y=center.coord[2]+R*(1-i*0.1), labels=strata.counts[i,1],c
  }
par(mfrow=c(1,1))
```



Example: Simulation for statistical efficiency) (stratification v.s. post stratification)

For more discussion, see this paper.

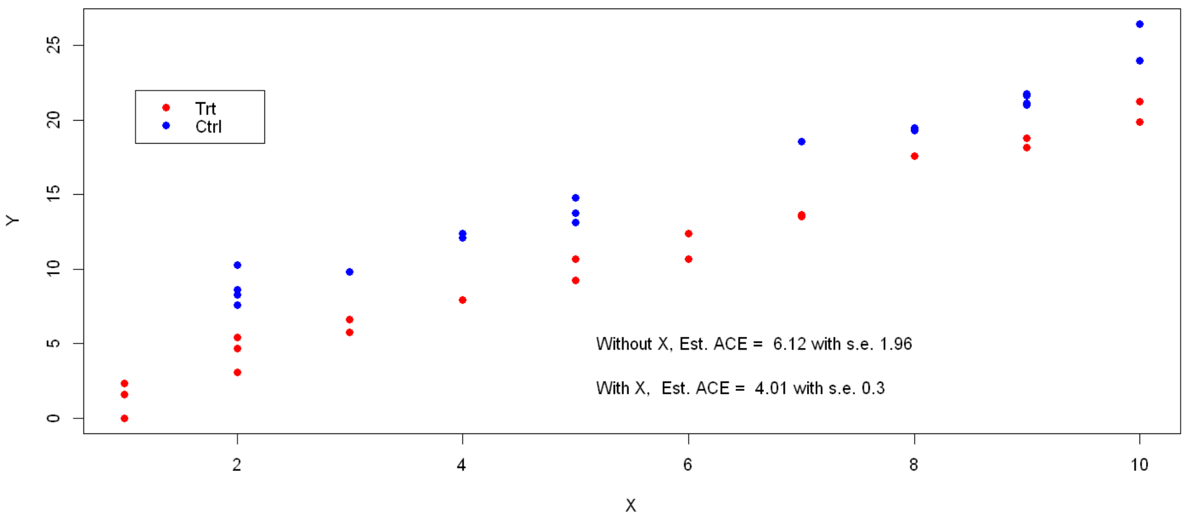
In [10⋯

```
# A simple simulation for stratification
```

```
set.seed(10928)
# Data generating mechanism:
n=40;n.strata=10;
X= sample(x=(1:n.strata),size=n,replace=TRUE);
ACE=4; coef.X=2;
Y.1=ACE+coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
Y.0=coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
trt= sample(1:n,size=(n/2),replace=FALSE);Z=rep(0,n);Z[trt]=1; # randomization

Z.s=rep(0,n);
for (i in 1:n.strata){# randomization within stratum
  id.stratum= which(X==i);
  trt= sample(id.stratum,size=floor(length(id.stratum)/2),replace=FALSE);
  Z.s[trt]=1;
}

Y=Y.1*Z+Y.0*(1-Z); # observation w/o stratification
Y.s=Y.1*Z.s+Y.0*(1-Z.s); # observation w stratification
```

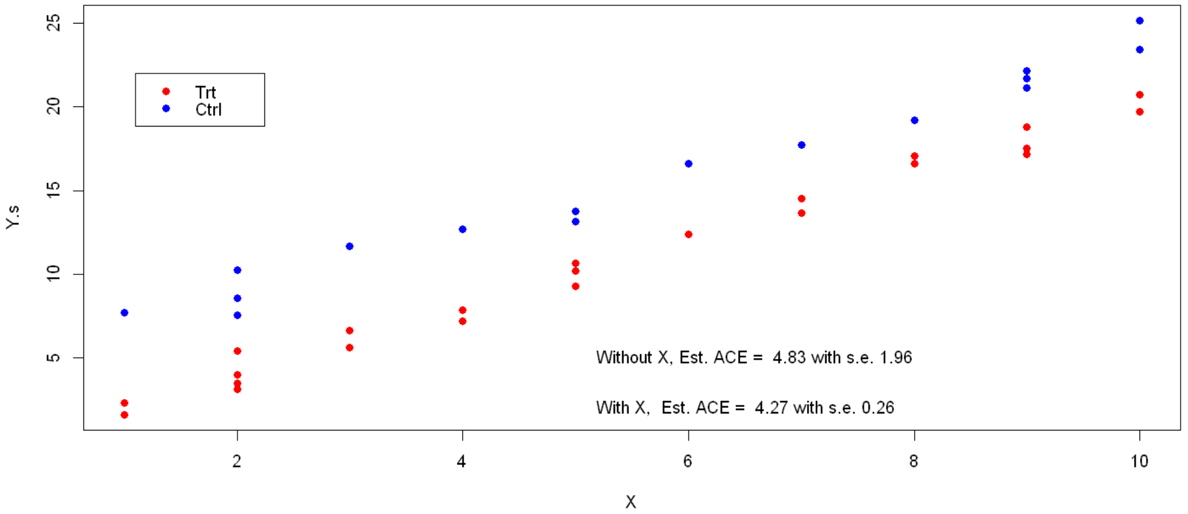
In [10⋯

```
# Analysis, w/o stratification

lm.vanilla=summary(lm(Y~Z));
lm.X=summary(lm(Y~Z+X));

options(repr.plot.width=12, repr.plot.height=6)
plot(y=Y,x=X,pch=16,col=c('red','blue')[Z+1])
legend(x=1.1,y=22,legend=c('Trt','Ctrl'),col=c('red','blue'),pch=16)
text(x=5.1,y=5,labels=paste('Without X, Est. ACE = ',round(lm.vanilla$coef[2,1],2), '

text(x=5.1,y=2,labels=paste('With X,  Est. ACE = ',round(lm.X$coef[2,1],2), 'with s.e
```



In [10⋯

```
# Analysis, w stratification

lm.simple=summary(lm(Y.s~Z.s));
lm.strat=summary(lm(Y.s~Z.s+X));

options(repr.plot.width=12, repr.plot.height=6)
plot(y=Y.s,x=X,pch=16,col=c('red','blue')[Z.s+1])
legend(x=1.1,y=22,legend=c('Trt','Ctrl'),col=c('red','blue'),pch=16)
text(x=5.1,y=5,labels=paste('Without X, Est. ACE = ',round(lm.simple$coef[2,1],2), 'w

text(x=5.1,y=2,labels=paste('With X,  Est. ACE = ',round(lm.strat$coef[2,1],2), 'with
```



In [10…

```
# Repeat the above procedure 10000 times to evaluate the efficiency

# Wrap up the code in one function

strat.sim<-function(ACE){
n=40;n.strata=10;
X= sample(x=(1:n.strata),size=n,replace=TRUE);
 coef.X=5;
Y.1=ACE+coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
Y.0=coef.X*X+rnorm(n,mean=0,sd=1); # potential outcome
trt= sample(1:n,size=(n/2),replace=FALSE);Z=rep(0,n);Z[trt]=1; # randomization

Z.s=rep(0,n);
for (i in 1:n.strata){# randomization within stratum
  id.stratum= which(X==i);
  trt= sample(id.stratum,size=floor(length(id.stratum)/2),replace=FALSE);
  Z.s[trt]=1;
}

Y=Y.1*Z+Y.0*(1-Z); # observation w/o stratification
Y.s=Y.1*Z.s+Y.0*(1-Z.s); # observation w stratification


lm.vanilla=summary(lm(Y~Z));
lm.X=summary(lm(Y~Z+X));

lm.simple=summary(lm(Y.s~Z.s));
lm.strat=summary(lm(Y.s~Z.s+X));

est.ACE=c(lm.vanilla$coef[2,1],lm.X$coef[2,1],lm.simple$coef[2,1],lm.strat$coef[2,1])
return(est.ACE)
}
ACE=4;

sim.result=replicate(n=1e4,strat.sim(ACE=ACE));
```

In [10…

```
mse=apply(sim.result-ACE,MARGIN=1,sd)
round(mse,digits=2)
```

4.59 · 0.32 · 1.36 · 0.32

10.3 Design of observational study

10.3.1 Overview

In Chapter 9, we mention that an observational study is a study where the treatments are not randomized. We have discussed some designs of randomized experiments, and one would expect that similar designs might apply to observational studies, where one needs to collect data.

There are other designs that also employ the stratified idea, for instance, repeated measures designs, split-plot design, nested design, etc. Note that the causal interpretation relies on randomization. A study can be stratified but not randomized. As a result, the causal interpretation might not always exist.

Depending on the length of observation, we would have the so-called longitudinal studies (or panel data) where subjects are observed over a period of time.

10.3.2 Reporting bias in survey sampling

In Chapter 9, we discuss how we can "remove" selection bias with assumptions. Here we demonstrate how to actually avoid bias with experimental design.

We consider the classic scenario of self-reporting bias. Responders of surveys are less likely to report true status if they worry about their privacies. In order to obtain true response, we can employ a randomized response technique (Warner model 1965). Let π_A be the sensitive proportion, i.e., proportion of subjects that belong of set A. With probability p, the question is whether you belong to set A (original question); with probability $1-p$, the question is whether you belong to set A^c (alternative question). Now considering the probability of saying Yes to the randomized question, we have

$$\text{pr}(\text{Yes}) = \text{pr}(\text{Yes} \mid \text{original question})\text{pr}(\text{original question}) + \text{pr}(\text{Yes} \mid \text{alt. question})\text{pr}(\text{alt.}$$

and

$$\text{pr}(\text{No}) = \text{pr}(\text{No} \mid \text{original question})\text{pr}(\text{original question}) + \text{pr}(\text{No} \mid \text{alt. question})\text{pr}(\text{alt. qu}$$

Hence we have $\hat{\pi}_A = [n_{\text{Yes}}/n - (1-p)]/(2p-1)$, which is unbiased as long as $p \neq 0.5$. In addition, we have

$$\text{var}(\hat{\pi}_A) = \frac{\pi_A(1-\pi_A)}{n} \frac{p(1-p)}{(2p-1)n}$$

10.3.3 Case-control study

Another popular design for observational study is the case-control study. In medical research, the cases are often easy to identify as patients' records are available from the health records. Furthermore, many diseases are rare in the general population. Therefore, rather than watching a group of participants till they develop diseases, it is more practical to select known cases and find matching control group.

Briefly, a case-control study identifies two populatons by the outcomes. The actual samples are drawn from the two populations, repectively. To be specific, the case group is drawn from $X, Y \mid Y = 1$, and the control group is drawn from $X, Y \mid Y = 0$. As a result, it is no longer possible to recover the ACE from a case-control study, as the averages are taken on wrong populaions.

However, the log-odds ratio from a case-control study can still be transfered to general population.