Contents

Chapter 1

Point and Interval Estimation

1.1 Z

$Z \sim N(0,1)$
$P(Z \leq z_p) = p$
$z_p = p^{th}$ quantile of standard normal distribution
$z_p = -z_1 - p$
Confidence interval for μ
$Y \pm z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$
To decrease width of CI: increase α or increase sample size
Assumptions:
Ys are sampled from a normal distribution
Variance is known
if σ^2 is not known estimate with s^2
However the distribution of $\frac{\overline{Y}-\mu}{s/\sqrt{n}}$ is not normal
Distribution of s^2
If an r.v. $Y \sim N(\mu, \sigma^2)$ then for a random sample size n, the quantity:

$$\frac{(n-1)s^2}{\sigma^2}$$

has a χ^2_{n-1} distribution

1.2 t Distribution

$Z \sim N(0,1)$ and $W \sim \chi^2_v$

If Z and W are independent then:

$$T = \frac{Z}{\sqrt{W/v}}$$

t distribution with v df
$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$
$100(1-\alpha)\%$ CI for μ:

$$\bar{Y} \pm t_{n-1,1-\alpha/2}\frac{s}{\sqrt{n}}$$

Ys are normal
If $n \geq 30$ w3 can use z as a reasonable approximation for t
Quantiles of t:
$qt(1-\alpha/2, df)$
CI
$t.test(t)\$conf.int$

1.3 Non-normal Data

If the Ys are not normally distributed use the CLT
CLT- If $Y_1, \ldots, Y_n$ is a random sample from a distribution with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ for $i = 1, \ldots, n$, then $\bar{Y}$ is approximately distributed as $N(\mu, \sigma^2/n)$ for large n
The use of CLT to construct a CI for μ requires knowledge of σ^2
Slutsky's Theorem when σ^2 is unknown
Slutsky's Theorem:
If X_n is a sequence of r.v.s that converges in distribution to X, and
Y_n is a sequence of r.v.s that converges in probability to a constant c,
Then $W_n = X_n Y_n$ converges in distribution to cX
That is:

$$\lim_{n\to\infty} P(W_n \leq w) = P(cX \leq w)$$

Let $X_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}}$ and $Y_n = \sqrt{\frac{\sigma^2}{s^2}}$
We know $X_n \xrightarrow{d} Z \sim N(0,1)$ and $\sigma^2/s^2 \xrightarrow{p} 1$
Then Slutsky's Theorem implies:

$$W_n = X_n Y_n = \frac{\bar{Y} - \mu}{\sigma}\sqrt{\frac{\sigma^2}{s^2}} = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

will be approximately $N(0,1)$
the approximation gets better as $n \to \infty$

Chapter 2

Hypothesis Testing

2.1 Critical Region and Value

Critical Region C_a
Boundaries of C_a are critical values
If the test statistic $S(\hat{\theta})$ is in C_a reject H_0
Otherwise fail to reject H_0

2.2 Tests of Hypotheses

1. Design Study
2. Establish null hypothesis
3. Determine test statistic to be employed
4. Choose α and establish C_a
5. Carry out study and collect data
6. Compute statistic from data
7. If statistic is in C_a reject H_0

Type I error: Reject H_0 when H_0 true
false positive
$\alpha = P(S(\hat{\theta}) \in C_\alpha | H_0)$
Type II error:
Do not reject H_0 when H_A true
false negative
$\beta = P(S(\hat{\theta}) \notin C_\alpha | H_0)$
Power $1 - \beta = P(S(\hat{\theta}) \in C_\alpha | H_A)$
Power is the probability of rejecting H_0 when H_A is true

p-value
Smallest significance level α for which the observed data indicate H_0 should be rejected.
Probability of obtaining a test statistic as unlikely or more unlikely than the observed test statistic if H_0 is true
p-value for 2-sided test
$2P(T < t)$
$2 * pt(t, df)$ one sided
$P(T < t)$

Chapter 3

One Sample Tests

p-value: Probability of obtaining a test statistic as extreme or more extreme than the one observed from the sample
μ_{diff}:
First subtract one pair from the other for a measurement then take the mean of the vector
$H_0 : \mu_{diff} = 0$
$H_A : \mu_{diff} \neq 0$
t-test assumptions for small sample:
observations are independent
sample is from the normal distribution
for large sample:
using CLT $Y \sim N(\mu, \frac{\sigma^2}{n})$
and Slutsky's Theorem $Z = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$
Ys don't need to be normally distributed
Can also test:
$H_0 : \mu = value$
$H_A : \mu \neq value$

3.1 Large Sample

Theorem: Reject H_0 if and only if CI excludes μ_0

3.2 Wilcoxon Signed Rank Test

Suppose $Y_1, Y_2, \ldots, Y_n$ are iid according to a symmetric distribution F with median $\zeta.5$
$H_0 : \zeta.5 = \zeta.5, 0$(median=specified value)
$H_A : \zeta.5 \neq \zeta.5, 0$

$wilcox.test(diff)$
Sign Test:
same hypothesis
$library("BSDA")$
$SIGN.test(diff)$
diff is the vector of subtracted pairs
signed rank test with ties:
$wilcox.test(diff, exact = F, correct = F)$

Chapter 4

Two Sample Tests

If two r.v.s Y_1 and Y_2 are independent then for any two constants a_1 and a_2 the r.v. $W = a_1Y_1 + a_2Y_2$ has mean and variance:
$E(W) = a_1E(Y_1) + a_2E(Y_2)$
$Var(W) = a_1^2Var(Y_1) + a_2^2Var(Y_2)$
Y_1 and Y_2 are independent and normally distributed then W is normally distributed with above mean and variance.
Corollary: If $\bar{Y_1}$ and $\bar{Y_2}$ are based on two independent random samples of size n_1 and n_2 from two normal distributions with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 then:

$$\bar{Y_1} - \bar{Y_2} \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

If the variances are equal then:

$$\frac{(\bar{Y_1} - \bar{Y_2}) - (\mu_1 - \mu_2)}{s_p\sqrt{1/n_1 + 1/n_2}} \sim t_{n_1+n_2-2}$$

where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

If $n_1 = n_2$ then:

$$s_p^2 = (s_1^2 + s_2^2)/2$$

4.1 Two sample t-test

$H_0 : \mu_1 = \mu_2$
$H_A : \mu_1 \neq \mu_2$
$C_\alpha = \{t : t > t_{1-\alpha,df}\}$
$t.test(data\$drug, data\$placebo, var.equal = T)$

4.2 Homogeneity of Variance

$H_0 : \sigma_1^2 = \sigma_2^2$ $H_A : \sigma_1^2 \neq \sigma_2^2$ If X_1 and X_2 are independent r.v.s with $X_1 \sim \chi^2_{v_1}$ and $X_2 \sim \chi^2_{v_2}$ then:
$\frac{X_1/v_1}{X_2/v_2} \sim F_{v_1,v_2}$
$qf(1-\alpha/2, df1, df2)$
$var.test(data\$drug, data\$placebo)$
(H_A : true ratio of variances is not equal to one)
If $\sigma_1^2 \neq \sigma_2^2$ and unknown:

4.3 Large Sample Approximation

If n_1 and n_2 are large, homogeneity of variance assumption is not important
Thus to test $H_0 : \mu_1 - \mu_2 = \delta$ we can use:

$$Z = \frac{(\hat{Y_1} - \hat{Y_2}) - \delta}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

Under $H_0, Z \sim N(0,1)$
Generally require $n_j \geq 25$ for $j = 1, 2$
Assumption that the Ys are normally distributed is no longer need because of CLT

4.4 Welch-Satterthwaite Approximation

Assume normality, n_1, n_2 small, $\sigma_1^2 \neq \sigma_2^2$
$t.test(data\$drug, data\$placebo, var.equal = F, alternative = "greater")$

4.5 Wilcoxon (Mann-Whitney) Rank Sum Test

Assume $Y_{1j}, \ldots, Y_{n_j j}$ iid $F_j(y)$ $j = 1, 2$
$H_0 : F_1(y) = F_2(y)$
$H_A : F_1(y + \Delta) = F_2(y)$
where Δ is a non-zero constant
$wilcox.test(data\$drug, data\$placebo, alternative = "greater", exact = F, correct = F)$
Wilcoxon Rank sum exact p-values:
$library(exactRankTests)$
$wilcox.exact(y1, y2)$
where y1,y2 are vectors
Rank Sum Test
$wilcox.test(data\$drug, data\$placebo, exact = F, correct = F, conf.int = T)$

4.6 Kolmogorov-Smirnov Test

$H_0 : F_1(y) = F_2(y)$ for all y
$H_A : F_1(y) \neq F_2(y)$ for at least one y
$D = max_y|F_{1n}(y) - F_{2m}(y)|$
where $F_{1n}(y)$ and $F_{2m}(y)$ are the EDFs for the samples F_1 and F_2
We reject H_0 only for large values of D
$ks.test(y1, y2)$

Chapter 5

Count Data

5.1 Comparing Two Proportions: Large Samples

If n_1 and n_2 are large we can use the normal distribution
Let n_{i1} be the number of successes in the i^{th} sample, $i = 1, 2$
Estimator of $\pi_1 = n_{i1}/n_i$
From CLT if n_i is large: $p_i \sim N\left(\pi_i, \frac{\pi_i(1-\pi_i)}{n_i}\right)$

2×2 table

	Success	Failure	
Sample 1	n_{11}	n_{12}	n_1
Sample 2	n_{21}	n_{22}	n_2
	m_1	m_2	N

Chi Square test
$H_0 : \pi_1 = \pi_2$
$H_A : \pi_1 \neq \pi_2$
$chisq.test(matrix(c(n_{11}, n_{12}, n_{21}, n_{22}), nrow = 2), correct = F)$
set correct=T for continuity correction

5.2 Measures of Association

In epidemiologic studies, we often obtain 2×2 tables

	Disease	No disease	
Exposed	n_{11}	n_{12}	n_1
Unexposed	n_{21}	n_{22}	n_2
	m_1	m_2	N

Estimands
$\pi_1 = P(diease|exposed)$
$\pi_2 = P(disease|not\ exposed)$
Risk Difference:
$RD = \pi_1 - \pi_2$
Risk ratio (relative risk):
$RR = \pi_1/\pi_2$
Odds ratio (cross product ratio):
$$OR = \frac{\pi_1(1-\pi_1)}{\pi_2(1-\pi_2)}$$
Independence or no association corresponds to: $RR = 1$ and $OR = 1$
$RR = 4$ implies an exposed person is 4 times as likely to have the disease as an unexposed person
$OR = 4$ implies the odds of disease in the exposed is 4 times that in the unexposed
$$OR/RR = \frac{1-\pi_2}{1-\pi_1}$$
If disease rare. $1 - \pi_1 \approx 1 - \pi_2 \approx 1$
Estimators
$\hat{RD} = p_1 - p_2 = (n_{11}/n_1) - (n_{21} - n_2)$
$\hat{RR} = p_1/p_2 = (n_{11}/n_1)/(n_{21} - n_2)$
$$\hat{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = \frac{n_{11}n_{22}}{n_{21}n_{12}}$$
In case-control studies $\hat{RR}$ should not be used to estimate RR
Measure of Association OR: $library(epitools)$
Rows should be the exposure, cols should be the cases status
Unexposed controls should be in the top left cell
$tab = array(table\ data, dim = c(2,2), dimnames)$
$oddsratio(tab)$

5.3 Confounding

A confounding variable is a variable that is associated with both the disease and the exposure
Can bias the measured association between exposure and disease
$H_0 : \pi_1 = \pi_2 \iff H_0 : OR = 1$
adjust for possible confounding by stratification and combining 2x2 tables
Mantel-Haenszel Test:
$H_0 : OR = 1$ within strata
create an array with dim $c(2, 2, 2)$
$mantelhaen.test(table)$

Chapter 6

Categorical Data

6.1 Contingency Tables

- Two-way $(r \times c)$ contingency table:

		j			
		1	2	$\cdots$	c
	1	n_{11}	n_{12}	$\cdots$	n_{1c}
	2	n_{21}	n_{22}	$\cdots$	n_{2c}
i	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	r	n_{r1}	n_{r2}	$\cdots$	n_{rc}

- Notation:

$$n_{i\cdot} = \sum_{j=1}^{c} n_{ij} \qquad n_{\cdot j} = \sum_{i=1}^{r} n_{ij}$$

Two scenarios where r x c tables arise:

1. Sample from a population and measure two characteristics say X and Y
   $P(X = i, Y = j) = \pi_{ij} \quad \sum_{i=1}^{r} \sum_{j=1}^{c} \pi_i j = 1$
2. Each row corresponds to a sample from a different population
   $\sum_{j=1}^{c} \pi_{ij} = 1$

χ^2 Test for Trend
$H_0 : \rho_1 = \rho_2 = \cdots = \rho_c$
$H_A : \rho_1 \leq \rho_2 \leq \cdots \leq \rho_c$
with at least one strict inequality
prop.trend.test()

Measure of Agreement κ
$library(vcd)$
$Kappa(table)$
$confint(Kappa(table))$

Chapter 7

Goodness of Fit Tests

7.1 Kolmogorov-Smirnov one sample test

We want to test whether our data came from a known and completely specified distribution: $F_0(y)$
$H_0 : Y_1, \ldots, Y_n \sim F_0(y)$
The KS statistic for GOF:
$D = max_y|F_0(y) - F_n(y)|$
or
$D = max\{D_1, \ldots, D_n\}$

7.2 Lilliefors KS GOF test

$library(\text{"}nortest\text{"})$
$lillie.test(x)$
KS test is preferred if data is continuous, KS test is more powerful in most situations than χ^2

7.3 Other GOF Tests

Shapiro-Wilk test for normality
$shapiro.test()$
Anderson-Darling, Crame-von Mises
$library(nortest)$ $ad.test()$ $cvm.test()$

Chapter 8

Regression I

Simple Linear Model
$Y = \alpha + \beta X$
Simple Linear Model with Error
$Y = \alpha + \beta X + \epsilon$
$\epsilon = Y - \alpha - \beta X$
ϵ is the vertical distance from Y to the line defined by $\alpha + \beta X$
Data are (Y_i, X_i) $i = 1, 2, \ldots, N$
Model Assumptions:

1. Linearity: $Y_i = \alpha + \beta X_i + \epsilon_i$
2. Xs are fixed constants
3. e_i iid $N(0, \sigma^2)$

Least Squares Estimation- LS estimators are values of α and β that minimize:
$\sum_{i=1}^{N} \epsilon_i^2 = \sum_{i=1}^{N} (Y_i - \alpha - \beta X_i)^2$
$\hat{\alpha} = \bar{Y} - \hat{\beta}\bar{X}$
$\hat{\beta} = \dfrac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2}$
If $X_i = Y_i$ for all i then $\hat{\beta} = 1$
If $Y_i = \bar{Y}$ for all i then $\hat{\beta} = 0$ Predicted Response:
$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$
Residual:
$r_i = Y_i - \hat{Y}_i$
Estimate Variance by MSE:
$\hat{\sigma}^2 = s_{y \cdot x}^2 = \dfrac{1}{N-2} \sum_i r_i^2$

Chapter 9

Rates and Proportions

Prevalence- proportion π of people with a particular disease at a fixed point in time
Rate- change in a variable over a specified time interval divided by the length of the time interval
Incidence - number of new cases of a disease in a period of time divided by the length of the time interval
Incidence is a rate, prevalence is not

9.1 Prevalence

N= random sample size from pop of interest
n= cases (have disease of interest)
Estimator of prevalence:
$\hat{p} = \frac{n}{N} = \frac{\text{number of cases}}{\text{number of controls}}$
CIs and tests for prevalence are based on $n \sim Binomial(N, \pi)$
where π is the pop prevalence

9.2 Incidence

Estimator of incidence:

$$\hat{I} = \frac{\text{number of new cases}}{(\text{sample size})X(\text{time interval})}$$

General form
$\hat{I} = c * \frac{\text{number of new cases}}{\text{sample size}}$
where $c = 1000/\text{time interval}$
we can uses binomial principals for CIs and tests.

9.3 Direct Standardization

Use to adjust rates/proportions for possible confounders
Three Step:

1. Divide samples into K categories of the potential confounder
2. Compute the proportion or rate in each confounder category
3. Compute the weighted average of confounder-specific proportions/rates

Choice of weights is based on standard or reference population
e.g. aggregate of samples in hand or governmental population survey

Chapter 10

Survival Analysis

Survival analysis: response is time to an event
Measure time from beginning of follow-up until an event such as disease, death or relapse
In clinical trial beginning is almost always time of randomization
In epi study this is usually time of initial exposure assessment
T^* denotes the possibly unknown survival time
Assume $T^* > 0$
Survival Function:
$S(t) = P(T^* > t) = 1 - P(T^* \leq t) = 1 - F(t)$
where $F(t)$ is the CDF of T^*
Properties:
$S(0) = 1 \quad S(\infty) = 0$
If $t_1 \leq t_2$ then $S(t_1) \geq S(t_2)$
Censoring:
Often we don't know the exact failure time of all subjects
Reasons for right censoring:
subject does not experience event of interest before end of study
subject is lost to follow up during study (withdraws,moves, dies from something else)
Failures can also be left or interval censored

- Let T_i^* and C_i denote the survival and right censoring times for the i^{th} individual
- Observe $T_i = \min\{T_i^*, C_i\}$
- Censoring indicator

$$\delta_i = \begin{cases} 1 & \text{if failure, i.e., } T_i = T_i^* \\ 0 & \text{if right censored, i.e., } T_i = C_i \end{cases}$$

- We observe (T_i, δ_i) for $i = 1, 2, \ldots, N$

Estimating $S(t)$:
No censoring- use 1-EDF
Otherwise Kaplan-Meier estimator
KM is a nonparametric maximum likelihood estimator (NPMLE)
Assumes independent censoring
Also know as the product limit estimator
If no censoring $KM = 1 - EDF$
Alternative: life table or actuarial method
$library(survivial)$
t=vector of times in order
delta= vector of 0 or 1 based on survival status
x= vector of ranks from 1 to n
$fit = survfit(Surv(t, delta) \sim x, conf.type = "plain)$
$summary(fit)$
$library(ggfortify)$
$autoplot(fit)$

10.1 Log Rank Test

Testing under minimal assumptions whether two survival functions are different
Use log rank when there is right censoring
Otherwise use Wilcoxon rank sum test
Log rank test
$H_0 : S_1(t) = S_2(t)$ for all t
Where $S_j(t) = P(T_j^* > t)$ for $j = 1, 2$
Let $t_{(1)}, t_{(2)}, \ldots, t_{(K)}$ be the distinct ordered failure times in the two groups combined

At each time $t_{(k)}$ construct the table:

Group	At risk	Events	Survive
1	$R_1(t_{(k)})$	m_{1k}	$R_1(t_{(k)}) - m_{1k}$
2	$R_2(t_{(k)})$	m_{2k}	$R_2(t_{(k)}) - m_{2k}$
	$R(t_{(k)})$	m_k	$R(t_{(k)}) - m_k$

$survdiff(Surv(t, delta) \sim rx)$

10.2 Cox/Proportional Hazards Model

The proportional hazards model is a linear model for the log of the hazard or, equivalently, a multiplicative model for the hazard. $\log(\lambda(t)) = \log(\lambda_0(t)) + \beta X$ or $\lambda(t) = \lambda_0(t)\exp(\beta X)$
$\lambda_0(t)$ is called the baseline hazard
Consider two values of X, x_1 and x_2 then:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{\lambda_0(t\exp(\beta x_1))}{\lambda_0(t\exp(\beta x_2))} = \frac{e^{\beta_{x_1}}}{e^{\beta_{x_2}}}$$

independent of t
assumption of independence of t needs to be checked
Let X be an indicator of being in one of two exposure or treatment groups, then if $x_1 = 1$ and $x_2 = 0$:

$$\frac{\lambda_1(t)}{\lambda_2(t)} = \frac{e^{\beta\cdot 1}}{e^{\beta\cdot 2}}$$

e^{β} is the hazard ratio comparing group 1 and 2
$summary(coxph(Surv(t, delta) \sim rx))$

Chapter 11

Survey Sampling I

Sampling Study- Selecting some part of a pop to be observed so that one may estimate something about the whole of the pop
Typically want to estimate total or mean
Observational- does not intentionally disturb pop (not experimental)
One does have control over how the sample is selected

11.1 Terminology

Population- The group of units (people) we are sampling and studying, assumed to be of known, finite size
Sampling Design- the strategy followed in selecting a sample from a population
Sampling unit- Unit designated for listing and selection in a sample survey (eg persons, dwellings, households, area units, pharmacies)
Sampling frame- List of sampling units from which a sample is drawn
Variable- some measurement taken on members of the sample (eg number of children ever born to a woman aged 15-49 years) sometimes call this the y-variable or x-variable
Selection probability- likelihood, over repeated applications of a sampling design, that a particular unit will be chosen for a sample.
Probability Sampling- sampling in which the design calls for using random methods to ultimately decide which units are chosen. Every unit has a known, non-zero selection probability.
Equal Probability Sampling- Probability in which all units in the pop have the same selection probability. aka self-weighted sampling or epsem (equal probability of selection method) sampling
Non-probability sampling- Sampling in which subjective judgment (usually by interviewers) is used to decide who is chosen in the sample. Selection probabilities cannot be determined. Difficult to determine if the sample is representative.
Unbiased estimator- an estimator which if repeated over all possible samples

that might be selected using the sampling design, would yield estimates which on average equal the parameter being estimated. (eg sample mean from an SRS is an unbiased estimator of the pop mean). aka design-unbiased. key idea is the randomness in the estimator is induced by the sampling design.
$library(survey)$

11.2 Simple Random Sampling

Let N denote the number of units in the pop.
aka sampling without replacement (SRSWOR) is the sampling design in which n distinct units are selected from the N units in the pop in such a way that every possible combination of the n units is equally likely to be the sample selected.
obtained through sequence of independent selections from whole pop each unit equally likely to be selected at each step, discarding repeat selections and continuing until n distinct units are obtained
$f \equiv n/N$ is the sampling rate or sampling fraction
Steps:

1. Number units in pop (ie sampling frame) from 1 to N
2. Select and record a random number between 1 and N
3. At each subsequent step, select a random integer between 1 and N. If it is the same as a previously selected number, discard it. Otherwise record it
4. Continue in this manner until n different numbers between 1 and N have been chosen
5. Population units corresponding to the selected numbers form an SRS of size n.

Alternative Approach:

1. Generate a random number from $U(0,1)$ for each unit in the pop
2. Sort in order of the random numbers
3. Take the first n units in the sorted list

Key Properties:
All possible SRS have the same chance of being selected
The prob that any one pop unit will be chosen is n/N
Selection probs in an SRS are not statistically independent
Estimating Population Mean
denote pop mean by:
$\mu = \frac{1}{N} \sum_{i=1}^{N} y_i$

Denote pop var by:
$\sigma^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \mu)^2$
Let Z_i indicate whether unit i is in the sample, that is $Z_i = 1$ if i is sampled, 0 otherwise.
The y_i are fixed, the Z_i are random
sample mean: $\bar{y} = \frac{1}{n}\sum_{i=1}^{N} y_i Z_i$
sample var: $s^2 = \frac{1}{n-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 Z_i$
The sample mean is unbiased: each Z_i is Bernoulli with $E(Z_i) = n/N$ thus:

$$E(\bar{y}) = \frac{1}{n}\sum_{i=1}^{N} y_i E(Z_i) = \frac{1}{N}\sum_{i=1}^{N} y_i = \mu$$

finite population correction factor (fpc)=$1 - \frac{n}{N} = \frac{N-n}{N} = 1 - f$
$library(survey)$
$data = df(y = c(yvals), fpc =)$
$design = svydesign(ids =\sim 1, data, fpc = \ fpc)$
$svymean(data\$y, design)$
$confint(svymean(data\$y, design))$
$svytotal(data\$y, design)$
$confint(svytotal(data\$y, design))$

11.3 Stratified Sampling

stratification- the process of dividing a population of units into distinct sub-pops called strata.
Strata are formed so that each pop unit is assigned only one strata
The pop is divided into H strata
let N_h denote the number of pop units in stratum h for $h = 1, \ldots, H$
The total number of units in the population is:
$N = \sum_{h=1}^{H} N_h$
let n_h denote the sample size for stratum h, so that the total sample size is:
$n = \sum_{h=1}^{H} n_h$
A sample size of n_h is selected by some prob design (eg SRS) from each of the H strata, independent of each other
Stratum-specific parameters (eg means, totals) are estimated separately using data from each of the H strata
An estimation of the pop parameter is produced by appropriately combining the H individual stratum estimates
If SRS is used within stratum, this is called stratified random sampling

11.3.1 Notation and Estimands

Let y_{hi} denote the variable of interest associated with unit i of stratum h
$(i = 1, \ldots, N_h);\ h = 1, \ldots H)$
Let $Z_{hi} = 1$ if the corresponding unit is in the sample, 0 otherwise
Stratum total:
$\tau_h \sum_{i=1}^{N_h} y_{hi}$
Population total:
$\tau = \sum_{h=1}^{H} \tau_h = \sum_{h=1}^{H} \sum_{i=1}^{N_h} y_{hi}$
Estimator of the pop mean:
$\bar{y} = \sum_h W_h \bar{y}_h$
where $\bar{y}_h$ is an estimator of the mean μ_h for stratum h
$E(\bar{y}_h) = \mu_h$ implies $E(\bar{y}) = \mu$
Estimator of the variance of $\bar{y}$
$\hat{Var}(\bar{y}) = \sum_h W_h^2 \hat{Var}(\bar{y}_h)$