معرفی قابلیتهای برنامه Stata

و مقایسه آن با سایر برنامه‌های آماری رایج

این برنامه یکی و شاید تنها برنامه آماری گسترده‌ای است که برای کارهای اپیدمیولوژیک تنظیم شده است. قابلیتهـای آن بسـیار گسـترده بوده و در بسیاری از شاخه‌های آماری دارای آخرین روشهای کشف شده‌است. در طول این دوره شما با توانمندیهای برنامه آشنا خواهید شد. ولی در اینجا برای درک بهتر برنامه، محدودیتهای آن ذکر می‌شود.

۱. قیمت زیاد برنامه و کتابهای راهنمای آن
۲. نیاز به دانستن دستورات و تایپ آنها، بر خلاف برنامه SPSS که عمده کارها با منوهـا انجـام مـی‌شـود در برنامـه stata شـما بایست اکثر دستورات را خودتان بنویسید.
۳. تفاوت ظاهر پنجره‌های برنامه با برنامه‌های تحت ویندوز
۴. محدودیتهای برنامه در انجام آنالیزهای بسیار پیچیده آماری مانند آنالیز Bayesian، و مـدلهای multistage hierarchal، البته در این خصوص از سایر برنامه‌های رایج مانند SPSS and SAS چندان کم ندارد ولی در مقابل برنامه‌هـای اختصاصـی همچون MLwin and WinBug تواناییهای کمتری دارد.
۵. توانایی کم برنامه در کار کردن در شبکه و update نمودن بانکهای اطلاعاتی بصورت online
۶. سخت بودن ورود اطلاعات، به همین دلیل کمپانی تولید کننده همزمان نرم افزار stata transfer را نیز تولید می‌کند که افراد به راحتی بتوانند اطلاعات خود را به برنامه وارد کنند.

برنامه stata در سه سطح به فروش می‌رسد که مشخصات آنها با یکدیگر متفاوت هستند. با استفاده از مطالبی که در کارگاه به شـما داده-شده‌است تفاوتهای آنها را بیان نمایید.

معرفی منوهای برنامه و نحوه باز و بستن پنجره‌ها

این برنامه چهار پنجره اصلی textbased و دو پنجره hyper text دارد. پنجره‌های متنی اصلی عبارتند از پنجره متغیرها، پنجره دستورات، پنجره نتایج و پنجره مرور دستورات. پنجره‌های hyper text عبارتند از پنجره search و پنجره خروجی نتایج که در قسمتهای بعد با آن آشنا خواهید شد.

البته پنجره‌های دیگری نیز در این برنامه وجود دارد که فرعی‌تر هستند و بعداً با مهمترین آنها که عبارتند از graph and do file آشنا خواهید شد.

کار عملی: لطفاً در برنامه stata پنجره‌های اصلی را شناسایی نموده و نحوه باز کردن، بستن، جابجایی و تغییر اندازه آنها را تمرین نمایید. آیا می‌توانید اندازه خط متن پنجره نتایج را تغییر دهید.

از منوی بالای برنامه استفاده نمایید و سعی کنید کاربرد گزینه‌های مربوط به منوی windows and Prefs را بیابید.

نحوه ورود اطلاعات از طریق برنامه و یا انتقال اطلاعات از سایر برنامه‌ها

همانگونه که بیان شد ورود اطلاعات در برنامه stata مشکل است. اما در صورت نیاز می‌توان از منوهای برنامه برای ساده‌تر نمودن کار استفاده کرد. در مجموع به سه شکل می‌توان اطلاعات را به برنامه وارد نمود که عبارتند از ۱. استفاده از برنامه stata transfer، ۲. ورود اطلاعات بصورت دستی ۳. خواندن فایلهای اطلاعاتی در برنامه به صورت مستقیم

کار عملی: با استفاده از help برنامه سعی نمایید کاربرد دستورات insheet, infile and input را بیابید.

آیا می‌توانید حدس بزنید که دستورات زیر چگونه عمل می‌کنند

```
. input str20 name age str6 sex
" ۱.     "A. Doyle" 22  male
" ۲.     "Mary Hope" 37 "female"
" ۳.     "Guy Fawkes" 48 male
 ۴.     end
```

با استفاده از منوهای بالای صفحه، گزینه edit را پیدا کرده و سعی نمایید با استفاده از آن دو متغیر sex و age را ساخته و ۳ رکورد را در آن وارد نمایید.

با استفاده از برنامه stata transfer یک فایل SPSS (anexaity.sav) را به stata تبدیل نمایید.

برای تبدیل فایل MS-excel نیاز به دقت بیشتری است. می‌توانید دلیل آن را بیان نمایید.

دستورات پایه

شروع کار با نرم افزار STATA 10

گردآوری و نگارش : مینا مهدویان

فایل births.dta : این فایل حاوی داده های ۵۰۰ مادری است که در یکی از بیمارستانهای بزرگ لندن بدنبال یک حاملگی تک‌قلویی در این بیمارستان زایمان کرده اند. لیست داده های گردآوری شده در جدول زیر نمایش داده شده اند.

Variable	Units or Coding	Type	Name
Identity number	–	categorical	id
Birth weight	grams	metric	bweight
Birth weight < 2500 g	1=yes, 0=no	categorical	lowbw
Gestational period	weeks	metric	gestwks
Gestational period < 37 weeks	1=yes, 0=no	categorical	preterm
Maternal age	years	metric	matage
Maternal hypertension	1=hypertensive, 0=normal	categorical	hyp
Sex of baby (numeric)	1=male, 2=female	categorical	sex
Sex of baby (alphabetic)	male, female	categorical	sexalph

نرم افزار STATA، یک زیرشاخه پیش فرض برای جستجو فایلهای داده‌ها دارد. دستور pwd را برای نمایش مسیر این زیرشاخه اجرا کنید.

pwd

برای تغییر زیرشاخه پیش فرض دستور زیر را اجرا کنید تا مسیر برداشت فایلها به زیرشاخه c:\data\kssc تغییر یابد

cd c:\data\kssc

برای بازکردن فایل داده های births.dta دستور زیر را اجرا کنید.

use births.dta

اگر مایل به بستن فایل births هستید، دستور clear را اجرا کنید.

برای مشاهده داده های وارد شده، دستورات زیر را می توانید اجرا کنید:

browse

edit

سوال۱. به نظر شما چه تفاوتی بین این دستورات وجود دارد؟

برای دیدن نام، نوع، شیوه نمایش و برچسب متغیرها دستور زیر را اجرا کنید:

describe

برای مشاهده اطلاعاتی بیشتر در مورد یک متغیر، کدهای استفاده شده برای ورود داده ها ، شروط/ کدها missing و توصیف آماری داده‌ها دستور زیر را اجرا کنید:

codebook gestwks , compact

سوال۲. با استفاده از دستور بالا، سن مادران در این بررسی را توصیف نمایید:

Mean:

SD:

Median:

دستور list

برای مشاهده داده های متغیر سن مادران این دستور را وارد کنید:

list matage

سوال۳. چه اتفاقی می افتد برای اینکه ادامه روند را متوقف کنید از چه دستوری باید استفاده کنید؟

با استفاده از این دستورات میتوانید تعداد مشاهدات را محدود کنید

list matage in 1/5

list matage bweight in 11/20

list in 1/5

list in 1/5,display

list in 1/5,table

سوال۴. سه دستور آخر چه تفاوتی با دستورات اولیه داشت؟

دستور frequency

از این دستور برای ساختن جدول فراوانی استفاده میشود دستورزیر را وارد کنید

tabulate hyp

tabulate hyp sex

tabulate hyp sex,row

tabulate hyp sex,col

سوال۵. میتوانید تفاوت دو دستور آخر را بیان و نتایج را تفسیر کنید؟

سوال۶. با استفاده از دستورات آخر یک جدول برای متغیرهای جنس نوزادان (sex) و تعداد هفته بارداری مادر کمتر از ۳۷ هفته (Preterm) رسم نمایید و تعداد missingرا مشخص کنید

سوال۷. آیا میتوانید با استفاده از دستور table میانگین و انحراف معیار دو متغیر جنسیت و وزن نوزادان هنگام تولد رسم کنید و تفاوت دو دستور table و tabulate را بیان کنید.

دستور if

```
list  bweight if  bweight <2000
```

سوال۸. میتوانید حدس بزنید این دستور چه کاری انجام میدهد؟

```
count if  bweight<=2000 & sex==1
```

سوال۹. با استفاده از دستور if تعداد نوزادانی که وزن آنها کمتر یا مساوی g۲۰۰۰ یا بزرگتر از ۴۰۰۰ باشد را محاسبه نمایید

دستور gen

با استفاده از این دستور می توان متغیر جدید ساخت

```
generate num1=1
browse
```

سوال۱۰. با استفاده از این دستور متغیر bweight را به متغیر bw تبدیل کنید

دستورLabel

```
label data "whatever you like"
describe
label var  gestwks "gestp"
describe
```

سوال۱۱. دستور اول و سوم چه تفاوتی با هم دارند؟

سوال۱۲. بعد از وارد کردن دستور سوم چه اتفاقی در برچسب متغیر رخ داد؟

دستور notes

برای اینکه به متغیرهای داخل فایل نوشته هایی اضافه کنید تا در آینده برای استفاده دوباره آنها را به خاطر بیاورید میتوانید از این دستور اضافه کنید

notes matage :maternal age in year

notes matage

دستور recode

به وسیله این دستور کد متغیرها تغییر میابد

recode sex 2=0,generate(sex2)

tabulate sex2 sex

با استفاده از دستور دوم میتوانید تغییر ایجاد شده را مشاهده کنید

سوال۱۳. آیا میتوانید بگویید بین دو متغیر sexو sex2 چه تفاوتی است؟

دستور summarize

sum

sum gestwks , de

sum bweight if hyp==0 , de

دستور آخر را تفسیر کنید

دستور egen

زمانی که یک متغیر مقادیر متعددی دارد باید آنها را گروه بندی کنیم و یک متغیر جدید درست کنیم

```
egen agegrp=cut( matage) ,at(20,30,35,40,45)

tab agegrp

egen agegrp=cut( matage) ,at(20,30,35,40,45) icodes
```

سوال۱۴.دستور شماره ۳ را وارد کنید چه پیغامی مشاهده میکنید؟

سوال۱۵.برای برطرف کردن این پیغام چه دستوری باید وارد کنید؟

سوال۱۶.دستور ۱و۳ چه تفاوتی با هم دارند؟

دستورduplicate

```
duplicates report

expand 3 in 1/2

duplicates drop
```

سوال۱۷.دستور دوم را تفسیر کنید

با استفاده از این دستور duplicate متغیر وزن هنگام تولد را مشخص کنید

دستورtabmore

دستورات tabulate که در بالا به آن اشاره شد، قابلیتهای محدودی دارند، خصوصا برای محاسبه محدوده اطمینان در زیر گروه ها. همچنین، آنها فقط درباره متغیرهای کیفی آنالیز توصیفی را انجام می دهند. یک دستور بسیار کاربردی که تقریبا همه پیامدهای مطالعات اپیدمیولوژیک (کیفی، کمی، میزان و ..) را پوشش می دهدف دستور tabmore است.

برای دیدن پنجره این دستور، فرمان زیر را اجرا کنید:

db tabmore

و با انتخاب گزینه های مناسب، برآورد میانگین و محدوده اطمینان ۹۵٪ را برای وزن زمان تولد به تفکیک وضعیت فشارخون مادر محاسبه کنید. صورت دستور اجرایی به صورت زیر خواهد بود:

tabmore, res(bweight) typ(metric) row(hyp) mean

نتیجه را تفسیر کنید

سوال۱۷.با استفاده از این دستور میانگین و محدوده اطمینان ۹۵٪ هفته حاملگی به تفکیک گروههای وزن پایین نوزاد محاسبه کنید

دستورeffects

برای محاسبه رابطه بین دو متغیر، شاخصهای اثر مختلفی قابل محاسبه هستند. می توان اختلاف میانگین را به عنوان اثر نهایی گزارش کرد و یا از نسبت شانس، نسبت خطر و یا خطر افزایش یافته استفاده کرد. دستور effects امکان محاسبه تمام حالات فوق را فراهم می کند. برای دیدن پنجره این دستور فرمان زیر را اجرا کنید:

db effects

مثلا، برای محاسبه اختلاف میانگین بین مادران فشارخونی با مادران سالم از نظر وزن زمان تولد نوزاد، دستور زیر را اجرا کنید.

effects,res(bweight)typ(metric)exp(hyp) exc md

نتیجه را تفسیر کنید

سوال۱۸.با استفاده از این دستور، OR رابطه بین فشار خون مادر و وزن کم هنگام تولد (lowbw) را بررسی نمایید.

log file

برای ذخیره کردن نتایج حاصل از آنالیز باید قبل از اجرای دستورات یک فایل متنی باز نماییم. در نگارشهای قبلی برنامه فایل خروجی فقط یک فرمت بنام log داشت ولی در نگارشهای جدید فرمت دیگری نیز به آن اضافه شده است که بنـام smcl اسـت. فرمـت جدیـد مخصـوص

برنامه stata است و نمی‌توان آن را در سایر برنامه‌ها مانند MS-word باز نمود ولی logfile به راحتی توسط سایر برنامه‌های ویرایشی متن قابل خواندن است. با استفاده از منوهای بالای صفحه یک logfile باز نمایید. سپس دستورات زیر را اجرا کرده و logfile را ببندید. دقت نمایید بعد از باز کردن و بستن logfile چه دستوری در صفحه مرور دستورات به صورت خودکار نوشته میشود.
logfile ذخیره شده را توسط برنامه wordpad و یا MS-word باز نمایید.

describe

summarize

tabulate hyp

do file

با استفاده از do file میتوان دستوراتی که در STATA استفاده میشوند را ذخیره کرد.در پنجره review دستوراتی را که اجرا کرده اید میتوانید به do fileتبدیل کنید. بعد از انجام این کار do fileخود را باز کرده و دستورات ذخیره شده را مشاهده کنید

مروری بر دستورات اولیه
و آشنایی با دستورات بیشتر

گردآوری و نگارش : علی اکبر حقدوست

در کارگاه دستورات زیر توضیح داده‌خواهد شد، سعی نماید بعد از شنیدن توضیحات با دستورات مذکور کار نموده و سپس در جاهای خالی در نظر گرفته‌شده، برای خود توضیحات لازم را یادداشت فرمایید تا در آینده بتوانید از دست نویس خود برای یادآوری مطالب کمک بگیرید.

Use

Clear

exit, clear

cd

pwd

save

view help <command>

F3

Ctrl+Break

Display

preserve and restore

edit

browse

page up

paytakht.in.70@gmail.com

count

www.danesh-isi.ir

ساختن متغیرهای جدید، دستورات gen and egen و recode

دستورات بسیار کاربردی برای ساختن متغیرهای جدید است. در ادامه با کاربردهای آن آشنا خواهید شد.

لطفا فایل anxiety را باز نماید.

متغیرها و محتویات فایل را به دقت بررسی نمایید.

آیا می‌توانید حدس بزنید که دستورات زیر چه کاری را انجام می‌دهند؟

```
gen score2=score
recode score2 min/8=0 8/max=1
```

به نظر شما افرادی که دقیقاً score هشت داشته‌اند در گروه صفر قرار گرفته‌اند یا در گروه ۱ ؟

برای پاسخ به این سوال از دستور زیر استفاده نمایید و سپس اطلاعات را بر اساس score مرتب نمایید. آیا مفهوم دستور زیر را کاملاً درک نموده‌اید

```
brow sco*
```

آیا بر اساس دستورات فوق می‌توانید متغیر جدیدی به نام age2 درست نمایید که افراد ۳۵ سال و یا بیشتر را از بقیه جدا نماید.

بنظر شما دستور drop age2 چه عملی انجام می‌دهد؟ امتحان کنید

لطفاً دستورات زیر را اجرا نمایید و نتیجه کار را با دستور بالا مقایسه نمایید. کاربرد این دستور چیست؟

```
recode age (min/34=1) (34/max=2), gen(age2)
```

```
gen age2=(age<=35)
```

توجه، ممکن است بعد از اجرای این دستور پیام خطا دریافت نمایید، با کمی دقت دلیل آن را خواهید یافت. راه حل آن نیز قبلاً گفته شده است.

```
gen age3=group(3)
egen age4=cut (age), at(20, 25, 30,40,50)
egen age5=cut (age), at(20(3)30,40,50)
```

با دقت فایل را بررسی نماید و سعی کنید مشخص کنید که دو دستور بالا چه عملی را انجام داده‌اند

متغیر جدیدی درست نمایید که در آن مشخص نمایید هر فرد، مبتلا به چند بیماری است، یعنی اگر فردی هم بـه anxiety and tension مبتلا بود (مقدار دو داشت)، فرد را مبتلا به هر دو بیماری دانسته و اگر هر دوی آنها یک بود، فرد را کاملاً سالم بداند و اگر هر یک از این متغیـر دو عدد ۲ داشت، فرد دارای یک بیماری شناخته‌شود.

به نظر شما دستور زیر چه عملی را انجام می‌دهد، بعد از اجرا سعی کنید با استفاده از دستور brow کاربرد این دستور را کشف نمایید.

```
egen disease=group(ten anx)
```

دستور زیر برای ساختن تفاوت بین میانگین score از امتیاز اخذ شده توسط هر فرد است. سعی نمایید ابتدا مفهوم آن را درک کرده و سپس آن را اجرا کنید.

```
egen scoave=mean(score)
gen deviate=scoave-score
```

یکی از محاسن بسیار بزرگ stata آن است که می‌تواند میانگین و یا سایر عملگرها را در زیر گروهها حساب نماید. بعنوان مثال اگـر بخواهیـد اختلاف میانگین اخذ شذه هر گروه جنسی را از امتیاز فرد کم نماییم می‌توانیم به شکل زیر عمل کنیم

```
sort sex
by sex: egen sexscav=mean(sore)
gen sexdevia=scoave-score
```

ذخیره ساختن نتایج تحلیلها و دستورات

دقت فرمایید برنامه stata اطلاعات فایلها را هم زمان با نتایج حاصل از آنالیز و همچنین دستورات یک جا ذخیره نمی‌کند. برای ذخیره نمودن فایلهای اطلاعاتی بایست به صورت دائم باید از دستور save استفاده نمایید. البته برای خیره موقت اطلاعات در حافظه جاری برنامه می‌توانید از دستورات preserve and restore استفاده نمایید.

دستور preserve را اجرا کنید و سپس دستور زیر را تایپ نمایید

```
drop age*
restore
```

از منوهای بالای پنجره استفاده نمایید و فایل تغییر یافته anxiety را ذخیره نمایید. آیا می‌توانید اسم فایل را موقع ذخیره کردن عوض نمایید.

برای ذخیره کردن نتایج حاصل از آنالیز باید قبل از اجرای دستورات یک فایل متنی باز نماییم. در نگارشهای قبلی برنامه فایل خروجی فقط یک فرمت بنام log داشت ولی در نگارشهای جدید فرمت دیگری نیز به آن اضافه شده است که بنام smcl است. فرمت جدید مخصوص برنامه stata است و نمی‌توان آن را در سایر برنامه‌ها مانند MS-word باز نمود ولی logfile به راحتی توسط سایر برنامه‌های ویرایشی متن قابل خواندن است.

با استفاده از منوهای بالای صفحه یک logfile باز نمایید. سپس دستور زیر را اجرا کرده و logfile را ببندید. دقت نمایید بعد از بـاز کـردن و بستن logfile چه دستوری در صفحه مرور دستورات به صورت خودکار نوشته‌می‌شود.

logfile ذخیره شده را توسط برنامه wordpad و یا MS-word باز نمایید.

مجدد همان مراحل فوق را انجام داده ولی این بار فرمت smcl را انتخاب نمایید.

سعی کنید با استفاده از دستور translate فایل smcl را به فرمت logfile تبدیل نمایید.

برای ذخیره اطلاعات مربوط به دستورات تایپ شده نیز می‌توان از دو روش استفاده نمود. روش اول استفاده از فایلهای دستوری اسـت کـه اصطلاحاً به آن do فایل گفته می‌شود و در روزهای آینده با آنها آشنا خواهید شد. روش دوم استفاده از دسـتور #review <number of last commands> است.

لطفاً دستور زیر را تایپ نمایید:

```
#review 20
```

تغییر نام متغیرها، جابجایی آنها و تعریف نمودن برچسب (label)

مسلماً یادآوری مفهوم متغیرها و مقادیر آنها بسیار مشکل است. برای حل این مشکل می‌توان برای متغیرها و مقادیر آنها برچسب ساخت.

دستورات زیر را اجرا نمایید

```
label var score "psychological score"
label define disease 1 "no" 2 "yes"
label value tension disease
label value anxiety disease
brow tension anxiety
```

سعی نمایید با درک مفهوم دستورات فوق، برای متغیرهای دیگر فایل برچسب درست نمایید.

آشنایی با دستورات summarize, list, describe

برای شناختن بهتر فایل اطلاعات بایست متغیرهای فایل و همچنین نوع متغیرها بررسی شوند. دستورات فوق این کمک را به ما می‌کنند. لطفاً دستورات زیر را اجرا نمایید

```
sum age

sum age*

sum score, detail

by trial: sum age
```

اگر این دستور اجرا نشد بایست ابتدا فایل رابر اساس متغیر مورد نظر (trail) مرتب نمود. برای این کار بایست از دستور sort استفاده نمود. برای آسانتر شدن کار می‌توانید به جای عبارت by در ابتدای دستور از bys استفاده نمود.

اگر در مقابل دستور sum هیچ متغیری نوشته نشود، چه اتفاقی می‌افتد.

list sex

list in 12/20

list in -10/l

دستور describe را با و بدون انتخاب نام متغیر اجرا نمایید. چه تفاوتی در نتایج آنها وجود دارد.

آشنایی با دستورات sort, memory, display

دستورات زیر را اجرا نمایید

---

dis 2*2

dis 15/_pi

آیا معنای pi_ را متوجه شدید. اگر متوجه نشدید می‌توانید عبارت فوق را بصورت زیر تایپ نماید.

dis _pi

sort age

---

sort age sex

دستورات فوق فایل را بصورت صعودی مرتب می‌کنند. آیا می‌توانید روشی را بیابید که فایل مذکور را به صورت نزولی بر اساس age مرتب نماید. اگر توانستید با دو خط دستور این کار را انجام دهید برای خود دست بزنید!

یکی از مشکلا ت بسیار رایج زمانی که شما با یک فایل بسیار بزرگ کار می‌کنید مشکل حافظه است. البته منظور حافظه رایانه و برنامه است نه حافظه خودتان! برای حل این مشکل شما می‌توانید فضای اختصاص یافته به برنامه را افزایش دهید. این کار البته ممکن است باعث کاهش سرعت کامپیوتر شود. در صورت نیاز توضیحات کاملتر در جلسه داده‌خواهد شد. برای آنکه مشخص نمایید که فضای اختصاص یافته به برنامه چقدر است می‌توانید از دستور زیر استفاده نمایید.

memory

از دستور describe نیز برای بررسی وضع حافظه می‌توان استفاده کرد. آیا کاربرد آن را در قسمت قبل دریافته‌اید.

برای تغییر میزان حافظه بایست از دستور زیر استفاده کرد

set memory # b/k/m/g

توضیحات دستور در کارگاه بیان می‌شود.

البته راه دیگری نیز وجود دارد و آن فشرده نمودن فایل اطلاعاتی است. برای این کار از دستور compress استفاده می‌شود. با استفاده از دستور describe حجم فایل جاری را محاسبه نمایید و سپس دستور compress اجرا و مجدد حجم آن را بسنجید. چه نتیجه‌ای می‌گیرید.

آشنایی با functions

عملگرها کمک می‌کنند تا بتوان آسانتر و دقیقتر محاسبات را انجام داد. به صورت کلی عملکرها به چند دسته اصلی تقسیم می‌شوند

۱. عملکرهای ریاضی مانند +، -،*،/، ^، pi_؛ این عملگرها عمدتاً برای محاسبات استفاده می‌شوند.

۲. عملگرهای متنی که بر روی متغیرهای متنی عملیلتی را انجام می‌دهند. مثلاً دو متن را با یکدیگر ترکیب می‌کنند (ترکیب نـام و نام خانوادگی افراد که در دو متغیر جدا ذخیره شده‌اند) و یا علائم زاید را حذف می‌کنند مـثلاً space هـای اضـافه را برمیدارنـد. همچنین می‌توان از آنها برای ساختن متغیرهای جدید بر اساس چند character اول و یا آخر یک متغیر استفاده نمود.

آیا شما برای آخرین عملگرهای متنی کاربرد می‌شناسید. کدهای مربوز به مثلاً روستاهای کشور را بیاد بیاورید.

---

۳. عملگرهای مقایسه‌ای مانند علائم علامتهای بزرگتر، کوچکتر و مساوی و یا نامساوی (=!)

۴. عملگرهای منطقی مانند برابر بودن که با دو علامت مساوی (==) نمایش داده می‌شود و علامت not (!)، or (|) و and (&)

حال با دانستن این عملگرها لطفاً تمرینات زیر را انجام دهید. لطفاً قبل از دیدن اجرا این فرامید، سعی نمایید مفهوم عبارت را درک و مقـدرا آن را پیش بینی نمایید.

```
disp sqrt(4)

list if age<35

list if age==35

list if age>35 & sex==2

list if (age>35 & sex==2) |(age<=30 &sex==1)

count if age!=35
```

در برنامه بعضی متغیرهای سیستم وجود دارند که نقش بسیار زیادی در تسهیل فرآیند کار ایفا می‌نمایند. به عنوان مثال مقدار _n برابر شماره رکورد است و _N برابر شماره آخرین رکورد در کل فایل و یا زیر گروه مورد نظر می‌باشد. برای درک بهتر مفهوم مثالهای زیر را انجام دهید.

```
gen id=_n

gen samsize=_N

bys sex: gen samsize2=_N
```

یکی از نقاط قوت برنامه آن است که شما می‌توانید به تک تک مقادیر موجود در خانه‌های یک فایل اطلاعاتی ارجاع نمایید. به عنوان مثال عبارت age[10] یعنی سن رکورد شماره ۱۰

برای روشن شدن موضوع اگر بخواهیم مقدار تغییرات score را از هر مرحله trial تا مرحله بعد در یک متغیر جدید ذخیره نماییم میتوانیم به شکل زیر عمل کنیم.

```
sort subject trial

by subject: gen diff=score-score[_n-1]

bro subject trial score diff
```

لطفاً بیان نمایید عبارت زیر چه عملی را انجام می‌دهد

```
bys subject: gen meansco=sum(score)/_n

by subject: replace meansco=meansco[_N]
```

بعد از اجرای هر دستور آماری، مقادیر گزارش شده در متغیرهای موقت در حافظه ذخیره می‌شود. بعنوان مثال بعد از اجرای دستور sum می-توان مقدار میانگین محاسبه شده در متغیری موقتی ذخیره می‌شود که با عبارت r(mean) شناخته‌می‌شود. برای درک بهتر موضوع لطفاً دستورات زیر را اجرا نمایید.

```
sum age

dis r(mean)

gen agedev=age-r(mean)
```

تمرین زیر انتخابی است

با استفاده از برنامه به سادگی می‌توان متغیرهای جدیدی ساخت و از برنامه خواست بر اساس قوانین مشخصی آنها را پر نماید. به عنوان مثال دستورات زیر از برنامه می‌خواهند که یک فایل با ۱۰۰ رکورد جدید ساخت و متغیر وزن را با میانگین ۷۵ و انحراف معیار ۱۰ و با توزیع نرمال ساخت

```
preserve
set obs 100
gen weight=uniform()
replace age=invnorm(weight)
.replace age=age*10+75
```

آشنایی با دستورات language syntax: if, by, weight, options

برای اجرای بهتر و دقیقتر دستورات این امکان وجود دارد تا آنالیز را به زیر گروه خاصی محدود نمود و یا برای هر یـک از رکوردهـا وزن در نظر گرفت.

به صورت کلی دستورات stata فرمت کلی زیر را دارند

by(s) <varlist>:commad <varlist> if….in…using…, options

مثال

bys sex: sum age if trail= =2, detail

توجه: بایست دقت نمود که برای if حتماً دو علامت = در کنار یکدیگر بایست تایپ شود (==).

همچنین از دستور weight می‌توان برای وزن دادن به رکوردها استفاده نمود.

```
Preserve
Clear
set obs 2
gen sex=1
recode sex 1=2 if _n= =2
gen number=_n*100
tab sex
tab sex [fweight=number]
```

البته می‌توان با استفاده از دستور expand نیز فایل را باز نموده و تحلیلها را در آن انجام داد.

```
expand number
list
restore
```

آشنایی با دستورات recode

دستور recode کمک می‌کند که مقادیر یک متغیر را تغییر و با قادیر جدید جایگزین نمود.

```
Preserve
recode sex 1=. in -40/l
```

```
tab sex
restore
recode sex (1=2) (2=1), gen(rsex)
 recode anx ten (1=2) (2=1), pre(r)
recode score (min/10=1 low) (10/15=2 intermediate) (15/max=3 high), gen(gscore)
```

آشنایی با دستور table و دستورات مشابه در توصیف متغیرها

در جزو داده شده در ابتدا کارگاه، به اندازه کافی در خصوص دستور tab توضیح داده‌شده‌است. در این قسمت سعی می‌شود با تمرینـات زیـر قابلیتهای مهم دستور tab به نمایش گذاشته‌شود. لذا با دقت ابتدا دستورات زیر را مطالعه و سعی نمایید مفهوم آنها را درک کنید و سپس آنهـا را اجرا فرمایید.

```
tab ten

tab ten anx

tab1 ten anx

tab sex, sum(score)

table sex, c(n trial mean age sd age med age min score)

bys sex:tab ten anx

tab ten sex, col

tab ten sex, row

tab ten sex, chi

tab ten sex, exact

table sex ten, c(mean age)

table sex ten, c(mean age) center

table sex ten, c(mean age) center row col

table sex ten, c(mean age) center row col format (%9.2f)

table sex ten anx

table sex ten anx, c(mean score)
```

```
tabstat score age

tabstat score age, by(sex)

bys ten: tabstat score age

tabstat score age, stat(mean sd min max) format(%3.1g)

tabstat score age, stat(mean sd min max)  col(stat) format(%3.1g)

tabstat score age, stat(mean sd min max cv q)  col(stat) format(%3.1g)

tab trial, plot

tab trial, gen(tr)

table ten anx [freq=sex]

tabi 30 20 \ 20 10
```

ترسیم نمودارهای مختلف و ذخیره سازی

گردآوری و نگارش : علی میرزازاده

لطفا فایل birhts.dta را باز کنید.

- **BOX PLOT**

برای ترسیم box plot وزن نوزاد هنگام تولد، دستور زیر را اجرا نمایید:

```
graph box bweight
```

تعداد زیادی از افراد زیر خط پایینی قرار گرفته اند. به نظر شما شکل توزیع چگونه است؟ این خط بر چه مبنایی ترسیم شده است.

با اجرای دستور زیر، توزیع وزن نوزادانی که از مادران مبتلا به فشارخون بالا متولد شده اند را با مادران نرمال مقایسه کنید:

```
graph box bweight, over(hyp)
```

این کار را به تفکیک نوزادان پسر و دختر انجام دهید:

```
graph box bweight, over(hyp) over( sexalph )
```

به جای over(sexalpha)، از by(sexalpha) استفاده کنید، و دو نمودار را با هم مقایسه کنید.

- **HISTOGRAM**

برای ترسیم هیستوگرام وزن زمان تولد، دستور زیر را اجرا کنید:

```
twoway histogram bweight
```

برای تغییر مبنای شروع محور x از صفر، پهنای هر ستون برابر با ۵۰۰ گرم و تبدیل محور y به درصد فراوانی، دستور بالا را به صورت زیر تغییر دهید:

```
twoway histogram bweight , percent start(0) width(500)
```

برای تمرین بیشتر، توزیع وزن نوزاد را به تفکیک بر اساس جنسیت نوزاد و وضعیت فشارخون مادر ترسیم کنید.

جالب است که در STATA می توان نمودار هیستوگرام را برای نمایش داده های کمی گسسته هم بکار برد، اما تنظیم آن کمی مشکل است.

```
twoway histogram hyp  , discrete xlabel(0 1) xscale(range(-1 2)) gap(50) percent
```

- **SCATTER PLOT**

برای نمایش نمودار نقاط پراکنده بین وزن زمان تولد نوزاد و هفته حاملگی که زایمان در آن اتفاق افتاده، دستور زیر را اجرا کنید:

```
twoway scatter  bweight  gestwks
```

به نظر شما، چه رابطه ای بین این دو وجود دارد؟

برای نشان دادن خط رگرسیون در نمودار فوق، باید از دو نمودار روی هم افتاده ، scatter و lfit، استفاده کنید:

```
twoway (scatter bweight gestwks) (lfit bweight gestwks) , xtitle("Gestation Perid, (weeks)")
ytitle("Birth Weight, gr")
```

حالت کلی دستور به صورت زیر است:

```
twoway ( … , options ) (… , options ) ,  options
```

آیا با استفاده از دستورات فوق، می توانید خط رگرسیون رابطه وزن زمان تولد نوزاد و هفته حاملگی را برای نوزادان پسر و دختر را جداگانه (اما در یک نمودار) ترسیم کنید؟

بسیار خوب، اجازه دهید اجزاء دستور زیر را با هم مرور کنیم

```
twoway (scatter bweight gestwks if sex==1, msymbol(x) color(black)) (lfit bweight gestwks if
sex==1 , color(black)) (scatter bweight gestwks if sex==2, msymbol(smcircle) color(red)) (lfit
bweight gestwks if sex==2, color(red)) , legend( label(1 "Boys") label(3 "Girls"))
```

- **Bar Chart**

برای ترسیم نمودار ستونی برای نمایش میانگین وزن زمان تولد به تفکیک جنسیت و ابتلا مادر به فشارخون بالا، دستور زیر را اجرا کنید:

```
graph bar (mean ) bweight, over(hyp) over(sexalph)
```

تصور کنید که وزن نوزادان در یک ماهگی هم اندازه گیری شود و با نام weight1 وجود داشته باشد، برای نمایش میانگین وزنها در کنار هم در یک نمودار ستونی به تفکیک جنسیت و ابتلا مادر به فشارخون بالا چه دستوری را پیشنهاد می کنید. برای مثال، به وزن نوزادان هر یک ۵۰۰ گرم افزوده و متغیر weight1 را بسازید، و سپس نمودار آنرا ترسیم کنید.

- **چند دستور تکمیلی**

دستور زیر را به خاطر دارید:

```
twoway scatter  bweight  gestwks  ,
```

به نظر شما اضافه کردن هر یک از دستورهای زیر به ادامه خطر بالا چه کاری را انجام می دهند :

```
title(TITLE)

subtitle(SUBTITLE)

caption(CAPTION)

note(NOTE)

by (sexalph, title(TITLE))

xtitle (Gestationa period in weeks, size(large))

ylabel(1000(500)5000)
```

برای ذخیره نمودار در حافظه موقت، از گزینه name(...) استفاده کنید:

```
scatter  bweight  gestwks , name(g1)
```

برای نمایش مجدد نمودار g1، دستور زیر را اجرا کنید:

```
graph  display g1
```

و برای ذخیره ان برای همیشه، از save استفاده کنید. مواظب حالت replace باشید:

```
graph  save g1 , replace
```

پسوند ذخیره شده gph است. برای ذخیره نمودار به فرمتهای دیگر ، مانند tif، از دستور زیر زمانی که پنجره نمایش نمودار باز است استفاده کنید:

```
graph export g1.tif , as(tif) replace
```

برخی از optionهای عمومی دستور graph را در جدول زیر مرور کنید:

Group	Option
Graph titles	`title(text, size( ))`
	`subtitle(text, size( ))`
	`caption(text, size( ))`
	`note(text, size( ))`
with by	`place the above options inside the by( )`
Axes	`xtitle(text, size( ))`
	`ytitle(text, size( ))`
	`xlabel(numlist, labsize( ) angle( ))`
	`ylabel(numlist, labsize( ) angle( ))`
	`xscale(range(numlist) log)`
	`yscale(range(numlist) log)`
Added lines	`xline(#, lpattern( ) lcolor( ))`
	`yline(#, lpattern( ) lcolor( ))`
Marker symbols	`msymbol( ) msize( ) mcolor( ) mlabel( )`
Connect style	`connect( )`
Legends	`legend(label(# "text") label(# "text") ...)`
	`legend(order(# "text" # "text") ...)`

paytakht.in.70@gmail.com

مبنای وزن‌دهی و آنالیز پیمایشهای کشوری با نمونه گیری خوشه‌ای یک‌مرحله ای

گردآوری و نگارش : علی میرزازاده

یک مثال ساده برای درک مفهوم وزن دهی:

برای برآورد درصد مصرف کنندگان سیگار در مردان کشور منالند، تعداد ۱۰ نفر با استفاده از روش نمونه گیری طبقه ای مورد مصاحبه قرار گرفتند. از هر استان (استان شمالی=۱، استان جنوبی=۲) ۵ نفر در سه گروه سنی (age_g=1,2,3) وارد مطالعه شدند. یک نفر در استان جنوبی در گروه سنی ۲ به سوال مربوط به سابقه مصرف سیگار پاسخی نداده است. جمعیت مردان کل کشور ۹۵۰ نفر (استان شمالی = ۵۵۰ مرد، استان جنوبی = ۴۰۰نفر) است. جمعیت هدف هر گروه سنی در هر استان در ستون pop ارائه شده است.

بدلیل اینکه نمونه گیری Proportional to Size نبوده است و میزان پاسخ دهی در همه گروهها یکسان نیست، لذا لازم است که برای دستیابی به بهترین (Unbais) برآورد از درصد مصرف سیگار، درصد وزن داده شده را محاسبه کنیم. جدول زیر را برای محاسبه مبنای وزن (برای هر گروه سنی به تفکیک استان) کامل کنید.

id	province	age_g	smk	Target pop. = N	None Missing = n	Weight = N/n
1	1	1	1	200		
2	1	1	1	200		
3	1	2	0	250		
4	1	2	0	250		
5	1	3	1	100		
6	2	1	1	250		
7	2	2	1	100		
8	2	2	.	100		
9	2	2	0	100		
10	2	3	1	50		
مجموع						

اکنون که خانه های جدول بالا را پرکردید، برای انجام محاسبات فوق در برنامه STATA ، فایل menaland.dta را باز کنید.

لطفا بدون اینکه کاغذ را برگردانید، با استفاده از دستور egen و تابع count تعداد افرادی که در هر گروه سنی در هر استان به سوال مربوط به مصرف سیگار پاسخ داده اند را محاسبه کنید.

سپس با تقسیم تعداد جمعیت هر گروه سنی در هر استان بر تعداد پاسخ دهندگان، متغیر weight را ایجاد نمایید.

[مگه نگفتم اول انجام دهید، بعد به این صفحه نگاه بندازید !!]

همانطور که شما هم متوجه شدید، دستورات زیر محاسبات بالا را انجام خواهند داد:

```
bys province age_g : egen non_miss = count(smk(

gen weight= pop/non_miss
```

با دستور list ، یکی بودن نتیجه محاسبات را با جدول صفحه قبل کنترل کنید. سپس برای محاسبه مجموع ستونها، دستور زیر را باید اجرا کنید:

```
tabstat pop non_miss weight  , stat (sum)
```

به نظر شما، آیا اعداد ستونها منطقی هستند؟ مفهوم هریک چیست؟

با محاسبه درصد وزن داده شده، پنجره one-way tables را با استفاده از دستور زیر باز کنید:

```
db tab
```

و با استفاده از پنجره weight و انتخاب مناسب متغیرهای لازم، محاسبه لازم را انجام دهید. نتیجه اجرای دستور زیر، درصد سیگار ۶۸/۴۲٪ است.

```
tabulate smk [aweight = weight]
```

تفاوت درصد وزن داده شده را از درصد اولیه، چقدر است؟ دلیل این تفاوت در چیست؟

به نظر شما آیا با توجه به روش نمونه گیری، آیا برآورد بالا بهترین برآورد است و یا باید نکات دیگر را مد نظر قرار داد؟

آنالیز داده های برنامه مراقبت غیر واگیر - پیمایش کشوری سال ۱۳۸۵

این پیمایش مطابق با روش پیشنهادی سازمان جهانی بهداشت (Stepwise approach) و با روش نمونه گیری خوشه ای یک مرحله ای در ۳۰ استان کشور (هر استان ۵۰ خوشه ۲۰ تایی) انجام شده است. فایل اصلی داده های این پروژه شامل سی هزار نفر و متغیرهای بسیاری است که تنها به انتخاب تعدادی از این افراد (۲۸ استان و در هر استان حدودا ۵ خوشه ۲۰ نفری) و تنها بخشی از متغیرهای آن اکتفا شده است.

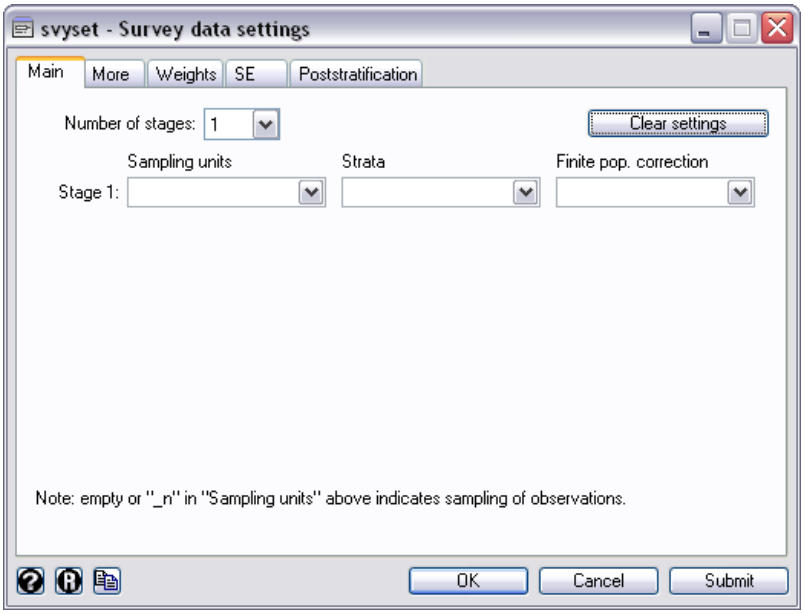
فایل STEPs85.dta را باز کنید. با اجرای دستورات زیر، مبنای وزن دهی داده ها، متغیر weight، را ایجاد نمایید.

```
bys id_pop : egen non_miss = count(s1)

gen weight= pop/non_miss
```

برای تنظیم ساختار داده ها بایستی مشخص کنید که کدام متغیر حاوی اطلاعات مربوط به کد طبقه، کدامیک مربوط به کد خوشه و کدام متغیر حاوی وزن داده ها ست. برای اینکار دستور زیر را اجرا کنید:

```
db svyset
```



در قسمت sampling units متغیر خوشه ها = partici2 ، در قسمت Strata متغیر طبقه ها = particip و در قسمت Sampling weight variable متغیر weight را جایگذاری کنید و کلید OK را فشار دهید :

```
svyset partici2 [pweight=weight], strata(particip) vce(linearized)
singleunit(missing)
```

پس از اجرای این دستور، دکمه save را فشار دهید تا برای همیشه ساختار نمونه گیری در این فایل داده ها ذخیره شود.

برای دیدن ساختار ایجاد شده، دستور svydes را اجرا نمایید:

svydescribe

آیا در خوشه ها گرد آوری شده، اشتباهی رخ داده است؟ کدام استان تعداد خوشه کافی را گرد آوری نکرده است؟

برای بررسی دقیقتر خوشه ها در هر طبقه، دستور زیر را اجرا کنید:

svydescribe , finalstage

برای بررسی دقیقتر در مورد یک متغیر، مانند s1، دستورات زیر را اجرا کنید:

svydescribe s1

svydescribe s1, finalstage

تمامی دستورات مربوط به survey analysis دز یک مجموعه واحد در قالب دستورات svy در STATA نمایش داده می شوند. در نسخه ۱۰ و بالاتر، دستورات svy به صورت حالت زیر تغییر ساختار داده اند :

Old Command	New Command
svygnbreg	svy: gnbreg
svyheckman	svy: heckman
svyheckprob	svy: heckprob
svyintreg	svy: intreg
svyivreg	svy: ivreg
svylogit	svy: logit
svymean	svy: mean
svymlogit	svy: mlogit
svynbreg	svy: nbreg
svyologit	svy: ologit
svyoprobit	svy: oprobit
svypoisson	svy: poisson
svyprobit	svy: probit
svyprop	svy: proportion
svyratio	svy: ratio
svyregress	svy: regress
svytab	svy: tabulate
svytotal	svy: total

برای محاسبه درصد فراوانی مصرف سیگار ، دستور زیر را اجرا کنید. پیشنهاد می شود از پنجره دستور برای اجرای استفاده کنید:

```
svy linearized : tabulate s1, se ci percent obs
```

اگر به تفکیک جنسیت، درصد فراوانی مصرف سیگار را لازم دارد، دستور زیر را اجرا کنید:

```
svy linearized : tabulate s1 c1, col se percent obs
```

دستور فوق بدون استانداردسازی انجام شده است. برای نمایش محدوده اطمینان ۹۵٪ درصد استاندارد شده مصرف سیگار به تفکیک جنسیت ، تغییرات زیر را در دستور بالا اعمال نموده و دوباره دستور را اجرا کنید:

```
svy linearized : tabulate s1 c1, stdize(id_std)
stdweight(std_pop) row percent ci
```

برای تمرین بیشتر در مورد کار با دستورات survey، ابتدا شاخص توده بدنی (bmi) را با استفاده از متغیرهای وزن و قد ساخته، سپس میانگین و محدوده اطمینان ۹۵٪ استاندارد شده آنرا را برای گروههای سنی به تفکیک جنس محاسبه نمایید.

سپس Adjusted OR را برای تاثیر متغیرهای سن، جنس و سیگار بر چاقی (bmi≥30) محاسبه کنید.

[برای اطمینان از جوابهای خود، svy_analysis.do را اجرا کنید.]

تمرین عملی

بررسی رابطه پلی مرفیسم NAT2 و GSTM1 با سرطان مثانه – مطالعه موردشاهدی جورشده

گردآوری و نگارش : علی میرزازاده

هدف اصلی این تمرین، آشنایی با مفاهیم آمار مقدماتی/ پیشرفته و انجام آنالیز آماری با نرم افزار STATA می باشد. در طول این تمرین اثر جورسازی بر نتایج، اثر انتخاب کنترلهای مختلف، اثر مخدوش کننده‌ها و اثر متقابل متغیرهای مختلف بر رابطه بین ژنهای مختلف با سرطان مثانه مورد توجه قرار می گیرد.

گروه مورد: شامل ۴۵۸ مورد جدید و قدیم سرطان مثانه تایید شده براساس یافته های بافتشناسی است. این موارد در مرکز ثبت سرطان ۱۲ بیمارستان در ۵ استان کشور اسپانیا بین سالهای ۸۶-۱۹۸۵ثبت شده بودند. بیمارستانهای واقع در ۲ استان بزرگ، جمعیت تحت پوشش کمی داشتند درحالیکه بیمارستانهای ۳ استان کوچک دیگر، درصد پوشش نسبتا خوبی، یعنی تقریبا برابر با جمعیت منطقه خود، را دارا بودند. تقریبا نیمی از موارد، موارد جدید بیماری بودند.

گروه کنترل: دو نوع کنترل، بیمارستانی و مبتنی بر جمعیت، در این مطالعه انتخاب شده اند. ۵۵۹ شاهد بیمارستانی از لیست بیماران پذیرش شده طوری انتخاب شدند که با گروه مورد براساس سن (±۵ سال)، جنس و محل زندگی یکسان باشند. بیمارانی که تشخیصی (هایی) مرتبط با عوامل خطر مورد مطالعه داشتند، از گروه کنترل خارج شدند. این تشخیصها شامل بیماری مزمن ریوی، بیماری قلبی، عفونت دستگاه ادراری،

هماچوری و سرطان راههای هوایی بودند. ۵۱۰ شاهد انتخاب شده از جامعه، از لیست شهروندی موجود در شهرداری طوری انتخاب شدند کـه از نظر سن و جنس با گروه مورد یکسان باشند.

مصاحبه : تمامی اطلاعات لازم با استفاده از یک پرسشنامه و در منزل افراد تکمیل شد

متغیرهای اصلی طرح :

در جدول صفحه بعد، متغیرهای مورد بررسی لیست شده اند.

سیگار: به صورتهای مختلف اندازه گیری شده است. Pack-years با ضرب تعداد سیگار در روز (Pack/day) در سالهای مصرف سیگار محاسبه شده است. برای ساخت متغیر "smkstat5" این متغیر در ۳۶۵ (روز) ضرب شده است.

پلی مرفیسم ژنی:

NAT2: در این تمرین، استیلاسیون سریع و متوسط در یک طبقه و استیلاسیون کند در یک طبقه گروه‌بندی شـده انـد. هفـت پلـی مرفیسـم (Single Nucleotide Polymorphisms – SNPs) مورد بررسی قرار گرفتند. افراد هموزیگـوت بـرای آللهـای استیلاسـیون سـریع (NAT2*4, NAT2*11A, NAT2*12B, NAT2*12C, NAT2*13) در گروه با فنوتیـپ استیلاسـیون سـریع قـرار گرفتنـد. افـراد هموزیگوت برای آللهای استیلاسیون کند، در گروه فنوتیپ استیلاسیون کند طبقه بندی شدند. افراد هتروزیگوت (یک آلل کند و یک آلل سریع NAT2) در گروه فنوتیپ استیلاسیون با سرعت متوسط جای گرفتند.

حذف GSTM1: در این تمرین، افراد بدون حذف یا هتروزیگوت با یک حذف، در گروه پایـه (+/+، +/– : reference group) و افـراد بـا دو حذف (–/–) در گروه مقابل جای گرفتند.

ساختار فایل داده ها

توضیحات مربوط به متغیرهای مختلفی که در این تحقیق گردآوری شده اند، را مشاهده می کنید:

variable	Codes	Description
id_num		patient ID number
case	0= control 1= case	Case control status
typ_co	1= case 2= Hospital based control 3= Community based control	Case control status: case/ Hospital based control / Community based control
gender	0= female 1= male	Self explanatory
age	continues	Self explanatory
ager	0= less than 59 years 1= 59-65 years 2= 66-71 years 3= more than 71 years	Age in 4 categories
smkstat2	0= never 1= ever	Smoking status : 2 categories
smkstat3	0= never 1= exsmoker 2= current	Smoking status : 3 categories
smkstat5	0= never 1= exsmoker 2= 1-10,000 pyrs 3= 10,001-15,000 pyrs 4= more than 15,000 pyrs	Smoking status in 5 categories
avg_ncig	continues	average number of cigarettes per day
numcigr	0= never smoke 1= less than 11 cig./day 2= 11-20 cig./day 3= more than 20 cig./day	avg_ncig in 4 categories
smktime	continues	Total duration of smoking in years
smktimer	0= never smoked 1= less than 30 years smoking 2= 30-40 years smoking 3= 41-48 years smoking 4= more than 48 years smoking	smktime recoded to 5 categories
packyrs	continues	Number of packyears (cumulative smoking exposure)

variable	Codes	Description
gstm1	0= “++/+-“ 1= “--"	GSTM1 Polymorphism
nat2	0= “Rapid/Intermediate” 1= “Slow”	NAT2 Polymorphism

نحوه اجرای تمرین:

هدف اصلی مطالعه بررسی رابطه پلی مرفیسم NAT2 و GSTM1 با سرطان مثانه است. علاوه براین، بدنبال بررسی اثر مخدوش کنندگی و اثر متقابل دیگر متغیرها بر این رابطه هستیم.

افراد در سه گروه جداگانه، داده ها را مورد تحلیل قرار داده و در روز پایانی گزارش آنالیز را ارائه می نمایند:

- گروه H: گروهی که تنها شاهدهای بیمارستانی را در نظر گرفته و آنالیز را تنها با انتخاب این گروه از شاهدها انجام می دهند.
- گروه P: گروهی که تنها شاهدهای مبتنی بر جامعه را در نظر گرفته و آنالیز را تنها با انتخاب این گروه از شاهدها انجام می دهند.
- گروه O: گروهی که بدون توجه به نوع شاهدها، آنالیز را بر روی داده ها انجام می دهند.

مراحل انجام کار به صورت زیر خواهد بود:

- آشنا شدن با فایل داده ها
- انجام Labelگذاری و Recoding لازم
- توصیف داده ها و گزارش Crude Analysis
- ارزیابی Confounding و Interaction
- ارزیابی احتمال وجود Selection bias
- بررسی Dose-response
- تفسیر یافته ها، نوشتن خلاصه مقاله
- ارائه روش آنالیز و نتایج بدست آمده

دستورکار

۱. **آشنایی با تمرین و سوال پژوهش:** در مورد نوع مطالعه، اطلاعات جمع آوری شده و اهداف اصلی آن بحث کنید. به نظر شما چه فرضیاتی را می توان برای این اهداف نوشت.

۲. **آشنایی با فایل داده ها:** فایل bladder.dta را باز کنید. دستورات describe و codebookرا اجرا کنید. با توجه به هدف اصلی، کدام متغیرها نقش وابسته، کدام متغیرها نقش مستقل و کدامیک ممکن است نقش مخدوش کنندگی داشته باشند؟ در مورد انتخاب بهترین متغیرها بحث کنید.

۳. **آماده سازی فایل داده ها:**

- متغیر case که مشخص کننده افراد مورد از شاهد است که به صورت 0 و 1 وارد شده است. دستور tab را اجرا کنید و خروجی را مشاهده کنید. دیدن 0 و 1 چندان جالب نیست. با استفاده از دستور label به روش زیر متغیر case را برچسب گذاری کنید:

```
label define case_la 0 control 1 case

label value case case_la
```

متغیرهای دیگر مانند typ_co، gender و ... را با روش بالا برچسب گذاری کنید.

[هرکجا که احساس کردید که کاملا دستور را یاد گرفتید، برای برچسب گذاری بقیه فایل lab_val.do را اجرا کنید.]

- متغیر age را براساس جدول متغیرها، با استفاده از دستور زیر گروه بندی نمایید:

```
recode age min/59=0 59/66=1 66/72=2 72/max=3, gen(ager)
```

کنترل کنید که گروه بندی را درست انجام داده باشید. سپس، همین گروه بندی را با دستور egen و گزینه cut انجام دهید.

برای تمرین بیشتر، متغیرهای smktimer و numcigrرا براساس گروه بندی ذکر شده در جدول متغیرها بسازید.

[اگر با این دو دستور آشنایی کامل دارید، تنها فایل recode.do را اجرا نمایید.]

٤. **انجام آنالیز توصیفی:**

- با استفاده از دستورات tabulate و summarize داده های این طرح را توصیف نمایید. برای مثال، دستور زیر را اجرا کنید و خروجی مربوطه را بررسی نمایید:

```
tab typ_co

tab case smkstat2, row m

bys case : sum age, de
```

این کار را برای توصیف بقیه متغیرهای طرح به تفکیک گروه مورد و شاهد انجام دهید.

برای اجرای دستورات بالا می توانید از پنجره آنها هم استفاده کنید. برای این کار دستور زیر را برای دسترسی به پنجره دستورtabulate اجرا کنید:

```
db tabulate
```

[اگر با دستورات فوق آشنا شده اید، برای مشاهده جواب این بخش، توصیف داده ها، describe.do را اجرا کنید.]

۵. انجام آنالیز خام بدون درنظر گرفتن همسان سازی:

- محاسبه OR خام خام **(Unadjusted - Crude Crude):** با استفاده از دستورات tabmore و effects رابطه بین tabulate و summarize داده های این طرح را توصیف نمایید. برای مثال، دستور زیر را اجرا کنید و خروجی مربوطه را بررسی نمایید:

```
tab case gstm1 , col

tabmore, res(case) typ(binary) row(gstm1) odds ci

mhodds case gstm1

effects, res(case) typ(binary) exp(gstm1) exc or
```

- با استفاده از دستورات بالا، ابتدا اثر پلی مرفیسم NAT2 را بر سرطان مثانه بررسی کنید و سپس به سوالات زیر پاسخ دهید:

  - چه مواجهه های دیگری با خطر ابتلا به سرطان مثانه رابطه دارند؟

- مفهوم OR برای هر یک از این مواجهه‌ها چیست؟
- مفهوم OR برای متغیرهای سن و جنس چیست؟

[اگر با دستورات فوق آشنا شده اید، برای مشاهده جواب این بخش، تحلیل خام، crude.do را اجرا کنید.]

٦. انجام آنالیز تطبیق یافته بدون درنظر گرفتن همسان سازی:

- در تمرین قبل رابطه بین GSTM1 و NAT2 با سرطان مثانه بررسی شد. مشاهده شد که این دو پلی مرفیسم با سرطان مثانه رابطه دارند. در مطالعات گذشته نیز اثر و مکانیسم این دو بررسی و اثبات شده است. علی رغم به نظر شما چرا بعضی از افراد که این دو ژنوتیپ را دارند (از نظر ژنتیکی مستعد بیماری هستند)، به بیماری سرطان مبتلا نشده اند؟
- در تمرین قبل نشان دادید که سیگار با سرطان مثانه رابطه دارد. برای بررسی تاثیر سیگار (متغیر سوم) بر رابطه بین ژن و سرطان مثانه، مسیر زیر را دنبال کنید.

effects, res(case) typ(binary) exp(gstm1) str(smkstat2) exc or

آیا سیگار Effect Modifier رابطه GSTM1 با سرطان مثانه است؟ آیا سیگار رابطه GSTM1 و سرطان مثانه را مخدوش کرده است؟

به جای smkstat2 از متغیر دیگری ،smkstat3، که سیگار را دقیق تر اندازه گیری کرده است، استفاده کنید. آیا در تفسیر شما فرقی کرد؟

با استفاده از دستور mhodds، آنالیز بالا را تکرار کنید. خروجی ها را با هم مقایسه کنید.

- رابطه پلی‌مرفیسم NAT2 با سرطان مثانه را با درنظر گرفتن نقش احتمالی سیگار بررسی کنید.

[اگر با دستورات فوق آشنا شده اید، برای مشاهده جواب این بخش، تحلیل خام، crudeum.do را اجرا کنید.]

۷. مدلسازی رابطه SNPs با خطر سرطان مثانه با درگرفتن تاثیر سایرمتغیرها و همسان سازی

- با استفاده از مدل لجستیک رابطه بین GSTM1 و سرطان مثانه را بررسی نمایید. دستور زیر را برای انجام این کار اجرا کنید:

```
logit case gstm1  gender age
```

مفهوم این ضرایب را با یکدیگر بحث کنید. آیا روشی را برای بهتر بیان کردن مفهوم این ضرایب می شناسید؟

به نظر شما چرا سن و جنس در مدل آورده شده اند؟ تفسیر ضرایب آنها چیست؟

- با استفاده از مدل لجستیک، رابطه بین مواجهه های دیگر با سرطان مثانه را بررسی کنید. نتایج را به صورت Crude OR بیان کنید.

- برای بررسی تاثیر سیگار (سه حالتی) بر رابطه بین GSTM1 و سرطان مثانه، مدل زیر را اجرا کنید:

```
xi : logistic case gstm1 i.smkstat3*gstm1 gender age , or
```

مفهوم این ضرایب را با یکدیگر بحث کنید.

- به نظر شما اگر کسی در حال حاضر سیگار مصرف کند و ژنوتیپ –/– = GSTM1 را داشته باشد، شانس چندبرابری برای ابتلا به سرطان مثانه را نسبت به فرد غیرسیگاری -/- GSTM1 خواهد داشت؟

- جدول زیر را کامل نمایید. راهنمایی : برای محاسبه خط آخر از دستور زیر lincom استفاده شده است:

```
lincom  nat2 +  _Ismkstat3_2 + _IsmkXnat2_2, or
```

جدول ۱ – بررسی رابطه NAT2 با سرطان مثانه به تفکیک وضعیت مصرف سیگار

متغیرهای پیشگویی کننده	Odds Ratio (CI 95%)
در طول زندگی سیگار مصرف نکرده است استیلاسیون سریع / متوسط استیلاسیون کند	1 ..... (-)
در گذشته سیگار مصرف می کرده اما سیگار را ترک کرده است استیلاسیون سریع / متوسط استیلاسیون کند	..... (-) ..... (-)
در حال حاضر سیگار مصرف می کند استیلاسیون سریع / متوسط استیلاسیون کند	..... (-) 4.41 (2.63 – 7.40)

- تفسیر شما از یافته های جدول بالا چیست؟ چه توجیه بیولوژیکی برای این یافته ها می توانید داشته باشید؟

۸. انجام آنالیزهای دیگر، دوز – پاسخ:

- رابطه بین متوسط تعداد مصرف سیگار در روز با سرطان مثانه را با مدل لجستیک در حالتهای زیر بررسی نمایید:
  - تعداد مصرف سیگار = avg_ncup
  - دسته بندی تعداد مصرف سیگار و ورود آن به مدل به صورت خطی = numcigr
  - دسته بندی تعداد مصرف سیگار و ورود آن به مدل به صورت طبقه ای (Categorical) = i.numcigr
- به نظر شما کدام مدل برای بررسی رابطه ژن NAT2 و تعداد سیگار با سرطان مثاله بهترین است؟ معیار شما برای انتخاب چیست؟

۹. بررسی اثر انتخاب کنترلهای مختلف برای نتایج:

- بر اساس بهترین مدل انتخاب شده در بخش هشتم، اثر کنترلهای مختلف را در تفسیر رابطه NAT2 بر سرطان مثانه بررسی نمایید.

۱۰. گزارش یافته ها و ارائه آن:

- نتیجه مراحل ۹ گانه بالا را در قالب یک خلاصه مقاله در ۲۵۰ کلمه بنویسید و یک نفر را از گروه خود برای ارائه گزارش کار گروه انتخاب نمایید.
- می توانید برای ارائه خود، از جدول و نمودار برای نمایش OR خام و تعدیل شده استفاده نمایید