BIOSTATS 640 – Introduction to R
Fall 2024

Suggestions for Data Management and Analysis

https://www.mathcoachscorner.com/2012/06/whats-your-plan-stan/

***** Make a COPY of this to use as YOUR OWN living document !! *****

1. Study Description

1. Study Description - Checklist

__1. Statement of Overall Research Goal

__2. Primary Specific Aims and Research Hypotheses

__3. Secondary/Exploratory Aims

__4. Study Design/Sampling Scheme

__5. Inclusion/Exclusion Criteria

__6. Sample Size(s)

__7. Study Variables

__1.1. Statement of Overall Research Goal

- Rationale - The overall research goal establishes the context and focus
- Aim for a clear and informative single sentence statement
- Example: (Source, with permission: Katherine Reeves) "The investigators propose to examine whether phthalate exposure is associated with bone mineral density (BMD) and major osteoporotic fracture (MOF) using data from the Osteoporotic Fractures in Men Study (MrOS), the Health, Aging, and Body Composition Study (Health ABC), and the Women's Health Initiative (WHI)."
- Example: (Source, with permission: Laura Vandenberg) "This proposal aims to determine how exposure to BPS alters lactation parameters in mice, and whether exposures have long-lasting effects on future lactational outcomes."
- Example: (Source, with permission: Jane Kent) " The unique focus of this proposal is on understanding the cellular and molecular mechanisms of the greater muscle fatigue during dynamic work in healthy older adults and those with mobility impairments. "

__1.2. Primary Specific Aims and Research Hypotheses

- Rationale - Specific aims are critical in deciding what type of modeling should be performed.

- Rationale - The research hypotheses, in addition to informing the choice of modeling approaches, are ***extremely*** helpful in deciding which data visualizations to produce.

- Example: (Source, with permission: Katherine Reeves) "Specific Aims: #1. To evaluate if phthalate biomarkers are associated with BMD, A) cross-sectionally, and B) longitudinally. We hypothesize inverse associations between urinary phthalate biomarkers and BMD. #2. To evaluate if phthalate biomarkers are associated with future, major, osteoporotic fracture risk. We hypothesize that women and men with higher levels of urinary phthalate biomarkers will have higher major osteoporotic fracture risk, independent of established risk factors for fracture. We will explore associations between phthalate biomarkers and site-specific MOF risk."

- Example: (Source, with permission: Laura Vandenberg) "Aim 3: To quantify the effects of BPS exposure on lactation in subsequent pregnancies. After pregnancy, the mammary gland retains parity-induced epithelial cells (PI-ECs, progenitors of lobuloalveolar cells), allowing it to respond more robustly to a second pregnancy [17]. *Our working hypothesis* is that BPS exposures during pregnancy and lactation diminish the ability of females to respond to subsequent pregnancies by decreasing the number of PI-ECs in the mammary gland. To test this hypothesis, we will quantify PI-ECs in mice exposed to BPS, evaluate methylation of genes implicated in pregnancy 'memory' [17, 18], and quantify lactation outcomes in mice in second pregnancies."

- Example: (Source, with permission: Jane Kent) "Aim 2: Mechanisms of fatigue: molecular and cellular alterations to contractile function. The molecular and cellular mechanisms of fatigue will be evaluated *in vitro* by single muscle fiber contractile function (within fiber-type comparisons) from biopsy samples (n=20 participants per group). *Hypothesis 2.1.* Reductions in single fiber force from baseline to "fatigue" conditions (high [Pi] and $[H^+]$) will be due to greater slowing of myosin-actin cross-bridge kinetics (rate of force production and myosin attachment time), demonstrating a consistent mechanism for fatigue across the groups. The possibility that older muscle shows greater sensitivity to changes in [Pi] and $[H^+]$ (i.e., greater "fatigue") will also be evaluated. *Hypothesis 2.2.* The single-fiber force-calcium relationship (PCa_{50}) at baseline and with "fatigue" will be lower in older muscle such that OI<OH<YH, consistent with the results of Hypothesis 1.4 and providing the first direct evidence of a role for altered calcium sensitivity in the fatigue of older human skeletal muscle."

__1.3. Secondary/Exploratory Aims

- Rationale - Secondary/exploratory data analyses are important as "hypothesis generating".
- Typically, these analyses do NOT involve formal tests of significance and the reporting of p-values.
- As the name suggests, analyses to address secondary/exploratory aims emphasize exploration and typically include visualizations and confidence interval estimation.
- Analyses to address secondary/exploratory aims are also performed to corroborate the findings from analyses to address the primary specific aims.
- Analyses to address secondary/exploratory aims may also be useful in the preparation of future grants.

- Example: (Source, with permission: Katherine Reeves) "Exploratory Aims: To explore: 1) potential mediation by BMD, and 2) potential effect modification by: A) biological sex, B) race, and C) exogenous hormone therapy."

- Example: (Source, with permission: Laura Vandenberg) "In secondary analyses, we will explore the nature and strength of apparent departures from the null. Of *a priori* interest here will be comparisons of each of the BPS dose groups relative to control and an exploration of apparent dose-response relationships. As appropriate to post-hoc analyses, we will limit our reporting of secondary analyses to confidence intervals and data visualizations, with no formal tests of statistical significance".

- Example: (Source, with permission: Jane Kent) "In addition to our primary analyses, we will perform univariate and multivariate regression analyses in both Aims, to explore associations between energetic variables and participant characteristics such as sex and PA metrics".

__1.4. Study Design/Sampling Scheme

- Rationale - The study design and sampling scheme are critical in deciding what modeling approach is appropriate.

- Rationale - The study design and sampling scheme are also critical to the appropriate handling of multiple predictors (e.g., is the focus on testing one primary predictor as in a randomized controlled trial or is the goal to screen a large number of potential predictors to identify risk factors)

- Retrospective vs prospective?

- Observational vs intervention?

- Mixed effects (fixed and random)?

- Nature of missing data (e.g., type of missingness, censoring, etc.)?

__1.5. Inclusion/Exclusion Criteria

- Rationale - The inclusion/exclusion criteria defined the generalizability of study findings (note - to be clear, loss to follow-up is also a factor!!!)

- Example: (Source, with permission: Katherine Reeves Eligible participants will: 1) be enrolled in the MrOS Sleep Study, Health ABC, or the WHI Bone Density Study, and 2) have provided ≥2 urine specimens during study participation. We will exclude participants who: 1) had a MOF prior to the last urine sample selected for this project, and/or 2) have insufficient urine volume available for laboratory analysis.

__1.6. Sample Size and Power

- Rationale - Sample size and power are critical in guiding multiple predictor regression. A rough rule of thumb is that you should have a sample size of 12 or so for every predictor in your model.

- Sample size, overall and by group?

- Did you do any sample size and/or power calculations?

- Important - Which question do you want to answer? There are 3 types of sample size/power calcualtions:

    - **Solve for n (Hypothesis Testing):** What sample size do I need to detect with power=FILLIN when performing a one-sided/two-sided type I error = FILLIN (usually .05) FILLIN test of Null: FILLIN.
    - **Solve for n (Confidence Interval Estimation):** What sample size do I need to estimate NAMEOFEFFECT with a FILLIN (usually 95%) CI that has desired confidence interval width = FILLIN
    - **Solve for minimum detectable: Given my study has n=FILLIN.** What is the minimum detectable (e.g. group difference, effect of treatment, etc) that I can detect with power=FILLIN when performing a one-sided/two-sided type I error = FILLIN (usually .05) FILLIN test of Null: FILLIN.

__1.7. Study Variables

- Consider listing them by category of meaning -
    - demographics
    - baseline
    - intervention/exposure
    - clinical course/exposure course (e.g., in hospital outcomes)
    - outcome

- Consider listing them according to their use in analysis -
    - outcome
    - predictors of interest
    - variables defining groups to explore effect modification
    - potential confounder
    - potential mediating variable

2. Basics of Data Management

2. Basics of Data Management - Checklist

__1. Create a data dictionary/coding manual

__2. Always work with a copy of the source dataset (and store the original safely elsewhere)

__3. Create a clean and "ready to analyze" data set
- Handle missing values explicitly
- Name variables
- Label discrete variable values
- Create new variables as needed

__4. Flow Chart of Analysis Sample Recruitment, Consent, and Cohort Maintenance

__2.1. Data Dictionary/Coding Manual

- Rationale - So useful!!! Need I say more?
- Tips. 1) variable names must NOT have spaces (unless you enclose in single quotes in R);
  2) consider all lower case;
  3) aim for self-explanatory where possible
- Tip. Consider including an additional column for "remarks/notes", etc.

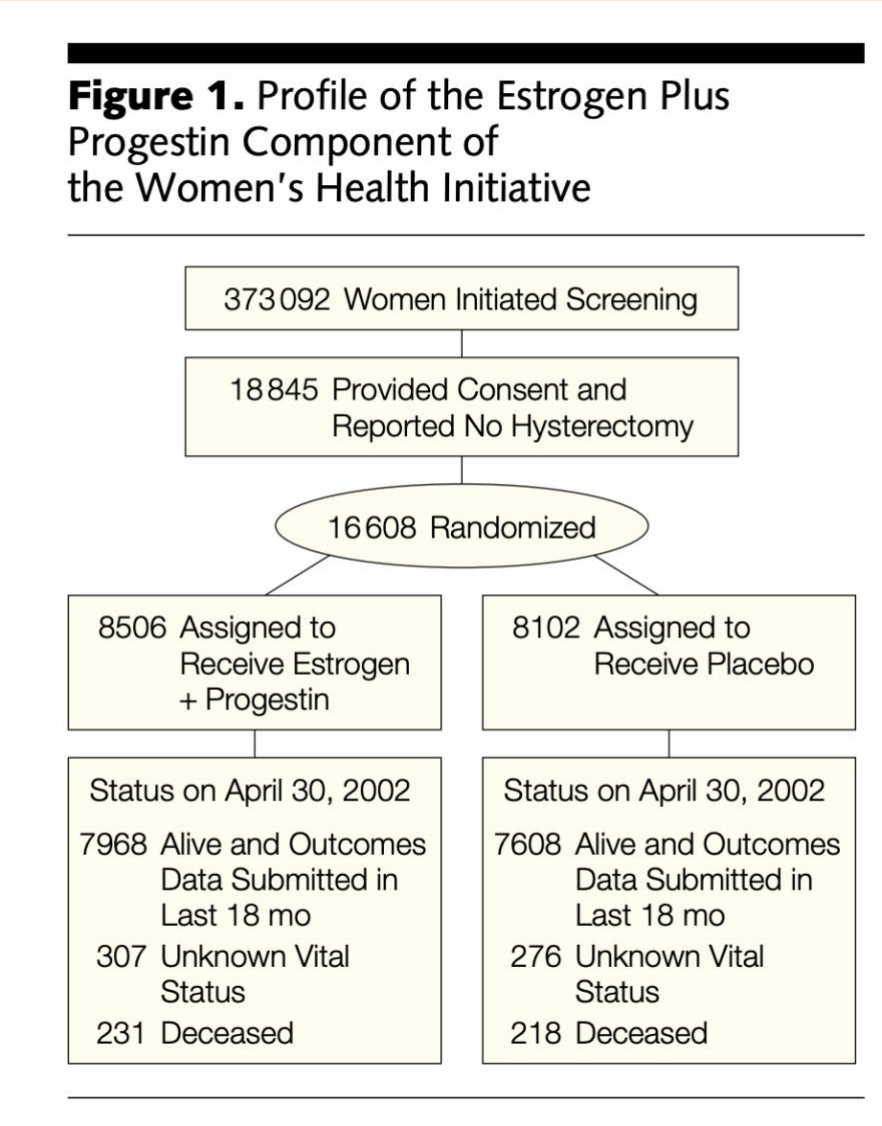
Data Dictionary/Coding Manual

Position	Variable	Variable Label (Units)	Type	Codes	Missing data
1	HT	Hormone Therapy	numeric	0 = placebo 1 = hormone therapy	None
2	age	Age in years	numeric	Range: [44, 79]	None
3	raceth	Race/ethnicity	character	"1" = White "2" = African American "3" = Other	None
4	exercise	Exercise at least 3x/week	numeric	0 = no 1 = yes	None
5	diabetes	Diabetes	numeric	0 = no 1 = yes	None
6	BMI	Body Mass Index, kg/m^2	numeric	Range: [15.49, 49.51]	Yes
7	glucose	Glucose, mg/dl	numeric	Range: [67, 294]	None
8	LDL	LDL Cholesterol, mg/dl	numeric	Range: [36.8, 365.2]	Yes

__2.2. Flow Chart of Analysis Sample Recruitment, Consent, and Cohort Maintenance

- Rationale - 1) Useful for assessing generalizability; 2) many journals require a figure depicting recruitment, consent, and cohort maintenance; and 3) this information is useful in future grant development
- Tip - Consider designing your own system for this work (e.g., an excel file or RedCAP instrument)
- Suggestion - Consider constructing a flow chart. The following is an example suitable for a randomized trials. Other study designs might call for a different flow chart schematic.

Example - Flow Chart



Source: https://jamanetwork.com/journals/jama/fullarticle/195120

3. Single Variable Distributions

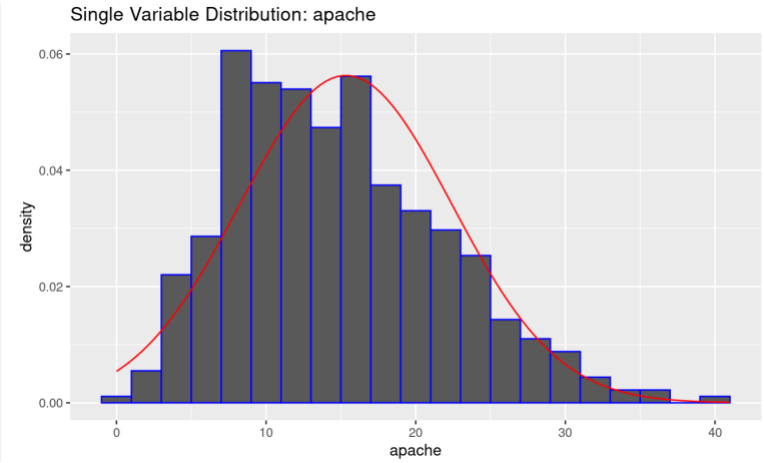
Encouraged!

- Create a document that contains a numerical and graphical summary of the distribution of EVERY variable.
- Organize these however works best for you (e.g., alphabetically by variable name, or alphabetically separately for categories of variables such as outcome, predictor, effect modifier, etc). Single page per variable?
- Good for: identifying errors, detecting missing values, range checks, visualizing patterns, outlier detection, discovering potential issues in model selection ... the list just goes on and on!

Example - Continuous Variable

Descriptive Statistics
sepsis$apache
Label: Baseline APACHE Score
N: 455

	apache
N.Valid	454.00
Mean	15.33
Std.Dev	7.09
Min	0.00
Q1	10.00
Median	14.00
Q3	20.00
Max	41.00
CV	0.46

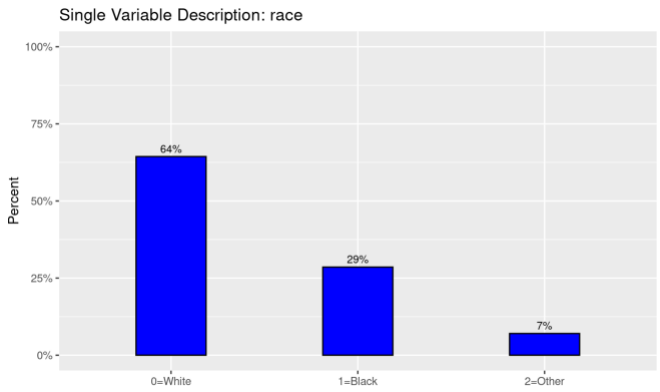



```
# numeric
library(summarytools)
descr(sepsis$apache,
      stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","CV"))
```

```
# graph
library(ggplot2)
ggplot(data=sepsis) +                                          # data = dataframe
      aes(x=apache) +                                          # x = single variable
      geom_histogram(binwidth=2,                               # you may have to play with this
                     colour="blue",                            # color = border of bars
                     aes(y=..density..)) +
 stat_function(fun=dnorm, color="red",                         # produce overlay normal, using:
               args=list(mean=mean(sepsis$apache,na.rm=TRUE),  # mean = mean of single variable
                         sd=sd(sepsis$apache,na.rm=TRUE))) +   # sd = sd of single variable
  ggtitle("Single Variable Distribution: apache")
```

Example - Discrete Variable

```
Frequencies
sepsis$racef
Type: Factor

               Freq   % Valid   % Valid Cum.   % Total   % Total Cum.
------------- ------ --------- -------------- --------- --------------
      0=White    293     64.40          64.40     64.40          64.40
      1=Black    130     28.57          92.97     28.57          92.97
      2=Other     32      7.03         100.00      7.03         100.00
         <NA>      0                                0.00         100.00
        Total    455    100.00         100.00    100.00         100.00
```




```
# numeric
library(summarytools)
freq(sepsis$racef, useNA="always")              # Tip. for data management, use option useNA="always"
```

```
# graph
library(ggplot2)
ggplot(data = sepsis) +
       aes(x = racef,                                                  # x = must be type factor
           y = prop.table(stat(count)),                                # optional: compute percents
           label = scales::percent(prop.table(stat(count)))) +         # optional: display as percents

       geom_bar(width=0.375, color="black", fill="blue",
                position="dodge",
                na.rm=TRUE) +

       geom_text(stat = 'count',
                position = position_dodge(.9),                         # these lines position labels
                vjust = -0.5,
                size = 3) +

       scale_y_continuous(labels = scales::percent, limits=c(0,1)) +   # limits=c(0,1) to set y-axis range

      ggtitle("Single Variable Description: race") +
      xlab(" ") +
      ylab("Percent")
```

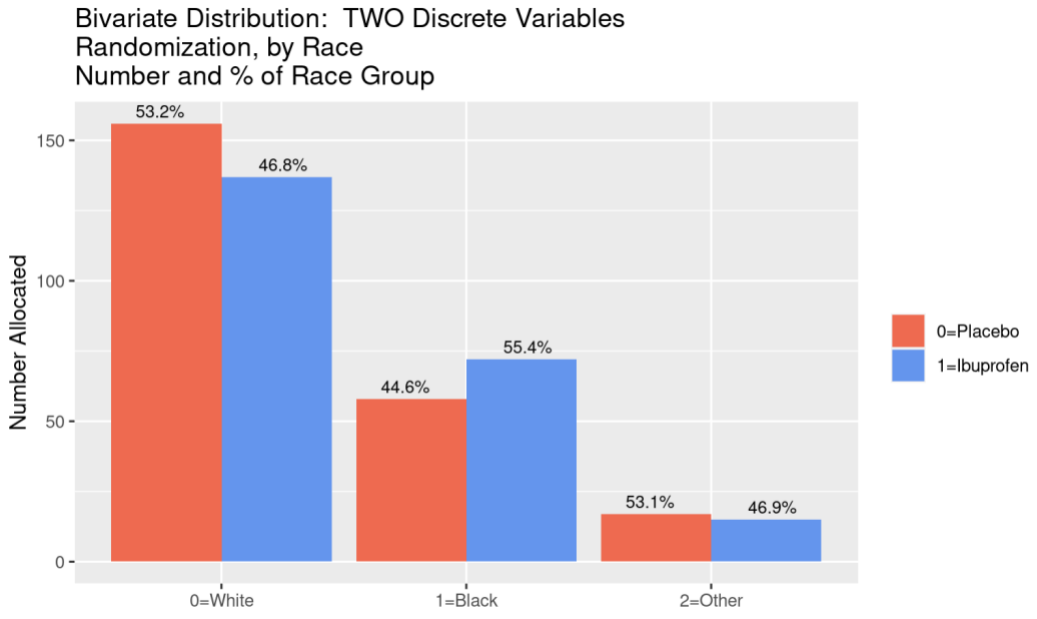
4. Bivariate (and Trivariate) Descriptions

So important! Do not ever start fitting models with out doing these data explorations first.

- Good for discovery of:
  - nature of relationship of outcome with predictors (e.g., linear v more complex);
  - patterns of association among the predictors themselves;
  - confounders;
  - effect modification.

Example - TWO discrete variables

```
Cross-Tabulation, Row Proportions
racef * treatf
Data Frame: sepsis

--------- -------- ------------- ------------- ---------- --------------
            treatf     0=Placebo   1=Ibuprofen       <NA>          Total
    racef
  0=White            156 (53.2%)   137 (46.8%)   0 (0.0%)   293 (100.0%)
  1=Black             58 (44.6%)    72 (55.4%)   0 (0.0%)   130 (100.0%)
  2=Other             17 (53.1%)    15 (46.9%)   0 (0.0%)    32 (100.0%)
     <NA>              0 ( 0.0%)     0 ( 0.0%)   0 (0.0%)     0 (  0.0%)
    Total            231 (50.8%)   224 (49.2%)   0 (0.0%)   455 (100.0%)
--------- -------- ------------- ------------- ---------- --------------
```



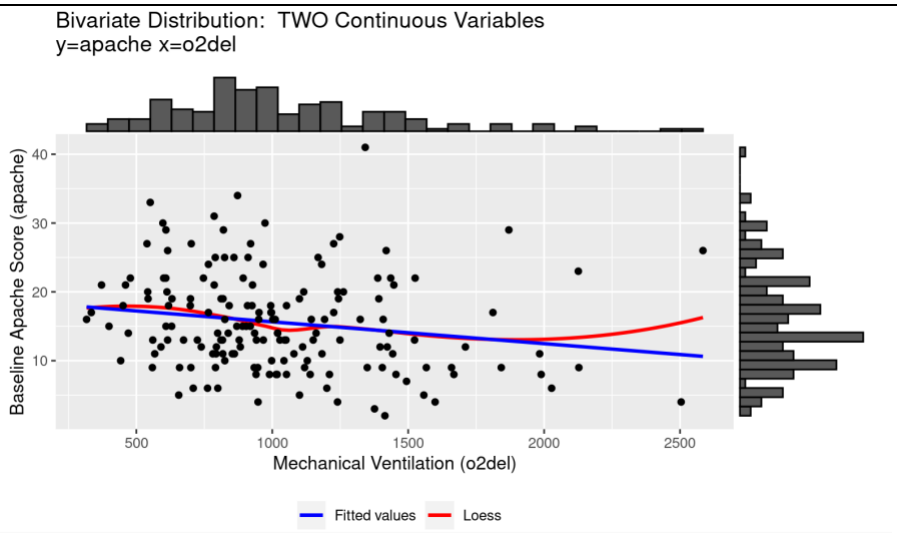

```
library(summarytools)
```

```
# create factor vars
sepsis$racef <- factor(sepsis$race,
                   levels=c(0,1,2),
                   labels=c("0=White", "1=Black", "2=Other"))
sepsis$treatf <- factor(sepsis$treat,
                   levels=c(0,1),
                   labels=c("0=Placebo", "1=Ibuprofen"))

# cross-tab
ctable(x = sepsis$racef, y = sepsis$treatf,prop="r",useNA="always")
```

```
# graph
library(tidyverse)
library(ggplot2)

plotdata <- sepsis%>%
      group_by(racef, treatf)%>%
      tally()%>%                                  # shorthand for summarise( ); easier to read
      mutate(percent=n/sum(n))                    # create percent = relative frequency

ggplot(data = plotdata,aes(x= racef, y = n,fill = treatf)) +         # x=predictor, fill=outcome

 geom_bar(stat="identity",                                           # show count
         position="dodge") +                                         # display side-by-side not stacked

 geom_text(aes(label=paste0(sprintf("%1.1f", percent*100),"%")),     # display % values on top of bars
               colour="black",                                       # color = color of text
               position=position_dodge(1.0),                         # these lines center the labeling
               vjust=-0.5,
               size=3) +

  scale_color_manual(values = c("coral2","cornflowerblue"))+         # optional:  choose your own borders
  scale_fill_manual(values = c("coral2","cornflowerblue")) +         # optional:  choose your own fill

  ggtitle("Bivariate Distribution:  TWO Discrete Variables\nRandomization, by Race\nNumber and % of Race
Group") +
  xlab("") +
  ylab("Number Allocated") +
  theme(legend.title=element_blank())
```

Example - TWO continuous variables

```
Descriptive Statistics
sepsis
N: 455

              N.Valid      Mean   Std.Dev      Min       Q1   Median        Q3       Max      CV
------------ --------- --------- --------- -------- -------- -------- --------- --------- ------
      apache    454.00     15.33      7.09     0.00    10.00    14.00     20.00     41.00   0.46
       o2del    168.00   1023.82    409.44   316.88   765.20   947.20   1233.08   2584.34   0.40
```

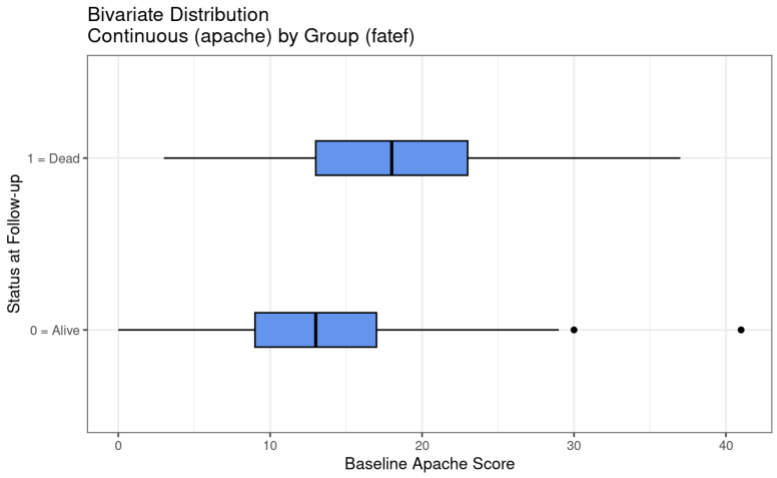



```
library(summarytools)
# numeric
descr(sepsis[c("o2del","apache")],
      stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","CV"),
      transpose=TRUE)
```

```
library(tidyverse)
library(ggplot2)
library(ggExtra)

# graph
plotdata <- sepsis %>%
            select(o2del,apache) %>%
            na.omit()                              # complete data only (n=168)

p <- ggplot(data=plotdata) +
       aes(x=o2del,y=apache) +
       geom_smooth(method = "loess",aes(color="Loess"), se=FALSE) +
       geom_smooth(method = "lm", aes(color="Fitted values"), se=FALSE) +
       geom_point() +
       scale_colour_manual(name="",values=c("blue", "red")) +
       ggtitle("Bivariate Distribution:  TWO Continuous Variables\ny=apache x=o2del") +
       labs(x="Mechanical Ventilation (o2del)",
            y="Baseline Apache Score (apache)") +
       theme(legend.position="bottom")
ggMarginal(p,type="histogram", size=5)        # main plot is size = 5x larger than marginal plot
```

Example - TWO variables: Continuous by Group

```
Descriptive Statistics
apache by fatef
Data Frame: sepsis
Label: Baseline APACHE Score
N: 279

                N.Valid    Mean   Std.Dev    Min      Q1   Median      Q3     Max     CV
--------------- --------- ------- --------- ------ ------- -------- ------- ------- ------
      0 = Alive    279.00   13.30      6.19   0.00    9.00    13.00   17.00   41.00   0.47
       1 = Dead    175.00   18.57      7.23   3.00   13.00    18.00   23.00   37.00   0.39
```




```
library(summarytools)
sepsis$fatef <- factor(sepsis$fate,
                       levels=c(0,1),
                       labels=c("0 = Alive", "1 = Dead"))
with(sepsis,                                                    # with(dataframe,
 stby(data = apache,                                               # data = continuous var
 INDICES = fatef,                                                  # INIDCES = group var
 FUN = descr,
 stats = c("n.valid","mean", "sd", "min","q1", "med", "q3", "max","CV"),
 transpose=TRUE))
```

```
library(ggplot2)

ggplot(data=sepsis)   +
 aes(x=fatef)         +                          # x = group var (must be factor)
 aes(y=apache)        +                          # y = continuous outcome

 geom_boxplot(width=0.2,                         # width of box
              color="black",                     # color = border
              fill="cornflowerblue") +           # fill =  box
 coord_flip()  +                                 # display boxes horizontally

 ggtitle("Bivariate Distribution\nContinuous (apache) by Group (fatef)") +
 xlab("Status at Follow-up") +
 ylab("Baseline Apache Score") +
 theme(legend.position = "none") +
 theme_bw()
```

5. Modeling to Address Primary Specific Aims

Plan Ahead!

__1. Modeling approach
- normal theory linear regression
- generalized linear model (logistic, poisson, etc.)
- time-to-event
- mixed-model

__2. What is the form of your primary outcome in data analysis?
- As is
- centered
- centered and scaled
- collapsed into groups
- transformed

__3. What summary statistic are you using to describe your primary association(s)?
- beta
- standardized beta
- relative odds (OR)
- relative hazard (HR)
- etc.

__4. What are your other variables and, for each, what is its role?
- primary predictors (continuous, discrete, 0/1, etc)
- effect modifier (usually defined using subgroups of some variable)
- confounder
- mediating

__5. Considerations in statistical hypothesis testing
- are you doing statistical hypotheses at all?
- null and alternative
- one-sided versus two-sided
- statistical burden of "proof"

__6. Suggestions for reporting; e.g.,
- table shell - descriptives of study participants
- table shell - "model free" estimates of associations of interest
- table shell - stratified "model free" estimates of associations
- additional table shells for results of modeling
- visualization of results of "model-free" and model results

6. Exploratory Analyses to Address Secondary Aims

Considerations

__1. Secondary/Exploratory analyses might include

- replication of main analyses in selected sub-groups
- replication of main analyses using alternative outcome variables
- assessment of dose-response
- identification of other predictors of outcome
- exploratory analyses to investgate "mechanisms" or "causality"

__2. Tools

- *NOT* formal statistical hypothesis tests and reporting of p-values
- data visualizations
- regression model development
- point and confidence interval estimation

7. Reporting

Example - Table Shell: Descriptives

Table 1. Baseline Characteristics of the Women's Health Initiative Estrogen Plus Progestin Trial Participants (N = 16 608) by Randomization Assignment* (cont)

Characteristics	Estrogen + Progestin (n = 8506)	Placebo (n = 8102)	*P* Value
Gail model 5-year risk of breast cancer, %			
<1	1290 (15.2)	1271 (15.7)	.64
1-<2	5384 (63.3)	5139 (63.4)	
2-<5	1751 (20.6)	1621 (20.0)	
≥5	81 (1.0)	71 (0.9)	
No. of falls in last 12 mo			
0	5168 (66.2)	5172 (67.5)	.18
1	1643 (21.0)	1545 (20.2)	
2	651 (8.3)	645 (8.4)	
≥3	349 (4.5)	303 (4.0)	

*Data are presented as number (percentage) of patients unless otherwise noted. BP indicates blood pressure; CABG/PTCA, coronary artery bypass graft/percutaneous transluminal coronary angioplasty; DVT, deep vein thrombosis; and PE, pulmonary embolism.
†Based on χ^2 tests (categorical variables) or *t* tests (continuous variables).
‡Required a 3-month washout prior to randomization.
§Total number of participants with data available was 8470 for estrogen plus progestin and 8050 for placebo.
||Among women who reported having a term pregnancy.
¶Statins are 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors.

source: https://jamanetwork.com/journals/jama/fullarticle/195120

Example - Table Shell: Regression Model Results

Tip. A good practice is to accompany reporting of OR and HR with number in group and # events in group

Table 2. Clinical Outcomes by Randomization Assignment*

Outcomes	No. of Patients (Annualized %) Estrogen + Progestin (n = 8506)	Placebo (n = 8102)	Hazard Ratio	Nominal 95% CI	Adjusted 95% CI
Follow-up time, mean (SD), mo	62.2 (16.1)	61.2 (15.0)	NA	NA	NA
Cardiovascular disease†					
CHD	164 (0.37)	122 (0.30)	1.29	1.02-1.63	0.85-1.97
CHD death	33 (0.07)	26 (0.06)	1.18	0.70-1.97	0.47-2.98
Nonfatal MI	133 (0.30)	96 (0.23)	1.32	1.02-1.72	0.82-2.13
CABG/PTCA	183 (0.42)	171 (0.41)	1.04	0.84-1.28	0.71-1.51
Stroke	127 (0.29)	85 (0.21)	1.41	1.07-1.85	0.86-2.31
Fatal	16 (0.04)	13 (0.03)	1.20	0.58-2.50	0.32-4.49
Nonfatal	94 (0.21)	59 (0.14)	1.50	1.08-2.08	0.83-2.70
Venous thromboembolic disease	151 (0.34)	67 (0.16)	2.11	1.58-2.82	1.26-3.55
Deep vein thrombosis	115 (0.26)	52 (0.13)	2.07	1.49-2.87	1.14-3.74
Pulmonary embolism	70 (0.16)	31 (0.08)	2.13	1.39-3.25	0.99-4.56
Total cardiovascular disease	694 (1.57)	546 (1.32)	1.22	1.09-1.36	1.00-1.49
Cancer					
Invasive breast	166 (0.38)	124 (0.30)	1.26	1.00-1.59	0.83-1.92
Endometrial	22 (0.05)	25 (0.06)	0.83	0.47-1.47	0.29-2.32
Colorectal	45 (0.10)	67 (0.16)	0.63	0.43-0.92	0.32-1.24
Total	502 (1.14)	458 (1.11)	1.03	0.90-1.17	0.86-1.22
Fractures					
Hip	44 (0.10)	62 (0.15)	0.66	0.45-0.98	0.33-1.33
Vertebral	41 (0.09)	60 (0.15)	0.66	0.44-0.98	0.32-1.34
Other osteoporotic‡	579 (1.31)	701 (1.70)	0.77	0.69-0.86	0.63-0.94
Total	650 (1.47)	788 (1.91)	0.76	0.69-0.85	0.63-0.92
Death					
Due to other causes	165 (0.37)	166 (0.40)	0.92	0.74-1.14	0.62-1.35
Total	231 (0.52)	218 (0.53)	0.98	0.82-1.18	0.70-1.37
Global index§	751 (1.70)	623 (1.51)	1.15	1.03-1.28	0.95-1.39

*CI indicates confidence interval; NA, not applicable; CHD, coronary heart disease; MI, myocardial infarction; CABG, coronary artery bypass grafting; and PTCA, percutaneous transluminal coronary angioplasty.
†CHD includes acute MI requiring hospitalization, silent MI determined from serial electrocardiograms, and coronary death. There were 8 silent MIs. Total cardiovascular disease is limited to events during hospitalization except venous thromboembolic disease reported after January 1, 2000.
‡Other osteoporotic fractures include all fractures other than chest/sternum, skull/face, fingers, toes, and cervical vertebrae, as well as hip and vertebral fractures reported separately.
§The global index represents the first event for each participant from among the following types: CHD, stroke, pulmonary embolism, breast cancer, endometrial cancer, colorectal cancer, hip fracture, and death due to other causes.

source: https://jamanetwork.com/journals/jama/fullarticle/195120

Template You Can Modify As You Like - Descriptives

Table XX Characteristics of Study Participants (n=511)

Variable	**Boys**		**Girls**	
Total, n	248		263	
Observations Parental BMI, n	231		222	
	n (%)		n (%)	
Ethnicity				
Caucasian				
Asian				
Black				
Chinese				
Other				
	mean (sd)	range	mean (sd)	range
Age, years				
BMI (kg/m^2)				
z-score				
FMI (kg/m^2)				
z-score				
LMI (kg/m^2)				
z-score				
Mother's BMI (kg/m^2)				
min/Q1/Q2/Q3/Q4/max				
Father's BMI (kg/m^2)				
min/Q1/Q2/Q3/Q4/max				

Key: sd – standard deviation, BMI – body mass index, FMI – fat mass index, LMI – lean mass index, Q2/Q3/Q4 - quintiles

Template You Can Modify As You Like - Results of Multiple Predictor Normal Theory Regressions

Table XX – Crude and Adjusted Estimated Change in Y=FILLIN (eg symptom score at baseline) (Beta, 95% CI) Associated with Increasing VALUE OF PRIMARY PREDICTOR (eg 0/1 metformin)

	Predicted Change Beta (95% CI) in Symptom score		
	Crude Beta (95% CI)	**Adjusted I Beta (95% CI)**	**Adjusted II Beta (95% CI)**
Primary Predictor			
Metformin (1=yes)			
Covariates			
Age, per 1 year?			
BMI			
Male gender			
Occupation			
Smoking			
Alcohol			
Creatinine			
More rows as needed			
Model Summary			
Adjusted R^2			
Model Comparisons[c]			
Model 2 v 1			
Model 3 v 2			

[a] 0/1, 1=FILLIN [b] Overall F test [c] 3 df Partial F test

Template You Can Modify As You Like - Bivariate Associations with 0/1 Outcome

Table XX - Crude Associations (OR, 95% CI) with Event of (Symptoms Score ≥ 4) (n=140, # events = fillin)

	N	# events (%)	OR	(95% CI)	Significance[b]
Continuous X, quantile			-	-	
Quantile 1					
Quantile 2					
Quantile 3					
Quantile 4					
Discrete X					
X=ref (eg female					
X=next level (male)					

[a] *Multiple logistic regression*

[b] *Likelihood ratio test on*

Template You Can Modify As You Like - Bivariate Associations with 0/1 Outcome, Stratified

Table XX - Crude Associations of Group with Ambulatory Care Visits (Event: 3 or More Visits in 12 Months)
N = 220, # events = 58 (26.4%)

	N	#events (row%)	OR	(95% CI)	P
TOTAL	220				
Releasees	58	20 (30.0)	1.98	1.02-3.84	.04
General	162	38 (22.9)	Referent		
Male	179				
Releasees	48	13 (27.1)	0.73	.34-1.55	.41
General	131	29 (22.1)	Referent		
Female	40				
Releasees	10	7 (70.0)	1.44	.66-3.09	.36
General	30	5 (17.7)	Referent		
Transgender	1				
Releasees	0	-	-	-	-
General	1	0 (0.00)	Referent		
Hispanic/Latino	61				
Releasees	13	6 (46.2)	0.93	.46-1.88	.84
General	48	8 (16.7)	Referent		
White	84				
Releasees	16	2 (12.5)	0.65	.33-1.23	.21
General	68	14 (20.6)	Referent		
Black or African-Am	60				
Releasees	25	10 (40.0)	1.46	.74-2.88	.28
General	35	9 (25.7)	Referent		
Other	8				
Releasees	2	1 (50.0)	1,94	.44-8.54	.38
General	6	2 (33.3)	Referent		
18-27 yrs of age	35				
Releasees	3	1 (33.3)	.68	.26-1.77	.43
General	32	5 (16.2)	Referent		
28-42 yrs of age	69				
Releasees	14	5 (35.7)	.96	.48-1.88	.90
General	55	11 (20.0)	Referent		
43-57 yrs of age	106				
Releasees	41	14 (34.1)	1.14	.60-2.18	.69
General	65	15 (23.1)	Referent		
≥58 yrs of age	10				
Releasees	0	-	-	-	-
General	10	3 (30.0)	Referent		
Housing-Stable/Perm	126				
Releasees	23	9 (39.1)	1.17	.61-2.23	.64
General	103	22 (21.4)	Referent		
AIDS Diagnosis	78				
Releasees	31	13 (41.9)	1.46	.77-2.80	.25
General	47	11 (23.4)	Referent		
Increased Need for Care	61				
Releasees	9	2 (22.2)	1.54	.77-3.07	.22
General	52	16 (30.8)	Referent		

[a] *1 df Likelihood ratio test*

Template You Can Modify As You Like - Crude and Adjusted Associations with 0/1 Outcome

Table XX - Crude and Adjusted Estimated Relative Odds (OR) of Event of Y=FILLIN (eg symptom score ≥ 4 at baseline) (Beta, 95% CI) Associated with Increasing VALUE OF PRIMARY PREDICTOR (eg 0/1 metformin)

	Predicted Relative Odds OR (95% CI) of Event of symptom score ≥ 4		
	Crude OR (95% CI)	**Adjusted I OR (95% CI)**	**Adjusted II OR (95% CI)**
Primary Predictor			
Metformin (1=yes)			
Covariates			
Age, per 1 year?			
BMI			
Male gender			
Occupation			
Smoking			
Alcohol			
Creatinine			
More rows as needed			
Model Summary[b]			
Model Significance)			
Model Comparisons[c]			
Model 2 v 1			
Model 3 v 2			

[a] 0/1, 1=FILLIN [b] Overall LR test [c] LR Test