In this lab, we’ll look at data from a study\(^1\) on nutrition, and the relationship between demographic variables, dietary variables, and plasma levels of beta carotene and retinol. The data were collected from people who had an elective surgical procedure to biopsy or remove a non-cancerous lesion of the lung, colon, breast, skin, ovary, or uterus.

You can get this dataset (and all other datasets from our textbook) here: [http://www.lock5stat.com/datapage.html](http://www.lock5stat.com/datapage.html). The data is called NutritionStudy, and it’s probably easiest to download the Minitab (.mtw) version.

There will not be prescriptive relationships for you to analyze, so feel free to choose whatever most interests you! Variables:

- **ID**: ID number for each subject in this sample
- **Age**: Subject’s age (in years) Smoke a factor with levels No Yes
- **Quetelet**: Weight/(Height\(^2\))
- **Vitamin**: Vitamin use: 1=Regularly, 2=Occasionally, or 3=No
- **Calories**: Number of calories consumed per day
- **Fat**: Grams of fat consumed per day
- **Fiber**: Grams of fiber consumed per day
- **Alcohol**: Number of alcoholic drinks consumed per week
- **Cholesterol**: Cholesterol consumed (mg per day)
- **BetaDiet**: Dietary beta-carotene consumed (mcg per day)
- **RetinolDiet**: Dietary retinol consumed (mcg per day)
- **BetaPlasma**: Plasma beta-carotene (ng/ml)
- **RetinolPlasma**: Plasma retinol (ng/ml)
- **Gender**: Coded as Female or Male
- **VitaminUse**: Coded as No Occasional Regular
- **PriorSmoke**: Smoking status: 1=Never, 2=Former, or 3=Current

1) **One Quantitative Variable**: Choose a quantitative variable, and use Minitab to explore it.

   a) Visualize the data in three different ways: histogram, dotplot, and boxplot.
      i) Discuss the shape of the distribution, and any other interesting characteristics.
      ii) Are there any obvious outliers?
      iii) What can you tell about the data just from the graph?
      iv) Which of the three types of graphs do you prefer for one quantitative variable?

   b) Calculate and discuss summary statistics (think about all the different statistics we talked about in class and try to interpret them from the output).

---

\(^1\) Nierenberg, Stukel, Baron, Dain, and Greenberg, "Determinants of plasma levels of beta-carotene and retinol", American Journal of Epidemiology (1989).
c) Based on both the graph(s) and statistics, describe this quantitative variable.

2) **One Quantitative by One Categorical:** Choose a quantitative variable and a categorical variable for which you are interested in analyzing the relationship, and use Minitab to explore the relationship.

   a) Visualize the data in three different ways: side-by-side boxplots, stacked dotplots, and overlaid histograms (and anything else you want to try). In all of these cases there is one Y (your quantitative variable) and you want to plot with groups (your categorical variable).
      i) Does there appear to be an association? Discuss the relationship.
      ii) What can you tell about the data just from the graph?
      iii) Which of these three types of graphs do you prefer for comparing groups?

   b) Calculate and discuss summary statistics for the quantitative variable by group. Which statistics (if any) are noticeably different? Which are the close to the same?

   c) Based on both the graph(s) and statistics, describe the relationship between this quantitative variable and this categorical variable.

3) **Two Quantitative Variables:** Choose two quantitative variables, and use Minitab to explore their relationship.

   a) Visualize the data with a scatterplot.
      i) Does there appear to be an association? Discuss the relationship. Does the association appear to be linear, nonlinear, or no association? If linear, is the association positive or negative?
      ii) Are there any noticeable outliers?
      iii) What can you tell about the data just from this graph?

   b) Calculate the correlation for these two variables (Stat -> Basic Statistics -> Correlation). What does this tell you about their association?

4) **Getting Fancier:** Beyond the basic scatterplot.

   a) Add a third variable (a categorical one) to your scatterplot by choosing the “With Groups” option from the scatterplot menu and adding a categorical grouping variable to the scatterplot. Discuss what this plot tells you.

   b) Add a third variable (a quantitative one) to your scatterplot by choosing “Bubble Plot” instead of Scatterplot – the size of the points will now be proportion to the third quantitative variable you choose. Discuss what this plot tells you.

   c) Combine (b) and (c) to add a third AND fourth variable (categorical and quantitative) to your scatterplot with “Bubble Plot” and “With Groups”. Discuss what this plot tells you.
d) Spice up the scatterplot from (3) (or make a new one if you wish), by choosing Graph → Marginal Plot instead of Scatterplot to add distributions of each of the quantitative variables individually to the plot. Discuss what this plot tells you.

e) Create and view simultaneously scatterplots for multiple pairs of quantitative variables. Choose "Matrix Plot" and enter as many quantitative variables as you want (more than two for it to be interesting). Discuss what this plot tells you.