Assignment 1

While the gap in gender inequalities has been shrinking over time, it no doubt still exists. In this question, we will examine salary discrimination amongst tenure‐track professors in a small Mid‐ western college in the United States. The data consists of information on 52 faculty members and were collected for presentation in legal proceedings for which discrimination against women in salary was at issue (recorded some time prior to 1980). The data were collected from personnel files, and consist of the following variables (note, some variables have been removed from the dataset we will work with). The data is saved in the file **salary.dta*.***

Provide your syntax (always, always use syntax to document your work!) and output. Round to three decimal places (unless otherwise stated) and assume a priori alpha for p<0.05.

Note on Documenting Syntax

Documenting syntax is critical!

* Not only will it enable you to record the record of your planned and unplanned analyses, but
* Well documented syntax will allow you to communicate your procedure to others (which is vital when you are working in teams or need to document your procedure for journal articles)

So, always provide a header (see example below) along with sufficient documentation for a reader to follow not just what you did, but WHY you did it.

See an example below:

Diagram

Description automatically generated

# Q1. Data Prep

Before we begin to analyze the data, we need to do a bit of cleanup. We have a few string variables. This isn’t a terrible problem, but Stata works much better with numeric variables. So, we’ll want to encode the string variables.

See https://www.stata.com/manuals14/dencode.pdf for tips on how to do this.

**Note:** we’ll want to use indicator coding (coding dichotomous variables using 0/1. Why? For this assignment, it really isn’t that important. However, in linear regression models, you will NEED to code dichotomous variables using indicator format since this will allow the models to generate accurate estimates of the coefficients. The reference category is typically coded as zero.

So, for instance, if we use the simple encode command for sex:

encode sex, gen(nsex)

and take a look at the data matrix, we will see that Stata generated a new numeric variable (nsex long %8.0g; wherelong %8.0g is Stata code for the type of numeric variable)

But, female = 1 and male = 2. This isn’t indicator coded like we’d like:

sex = recoded as “male=1” and “female=0”

So, we can use a nifty trick to tell Stata to indicator code (rather than let Stata make up codes for us).

label define n2sex 0 "female"  
encode sex, gen(n2sex)

**What are we doing here?** We are first defining the value of female in the label define command (yep, we are creating the label before we create the variable). Stata will then follow our lead and code female = 0 and male = 1 when we encode the new variable (n2sex).

Now, do the same for degree, and rank (note: because rank has three values, you can’t use the above shortcut; you’ll need to recode a different way).

degree = the highest degree obtained, recorded as “doctorate=0” and “masters=1”

rank = academic rank, recorded as “full=0”, “associate=1” and “assistant=2”

We don’t need to recode yearsdeg or salary

yearsdeg = the number of years since the degree was earned

salary = academic salary in U.S. dollars

### Q1.1 Once you have recoded your data using the approach described above, paste your syntax here.

# Q2. Data analysis

We will work through an analysis of this data. Include any of the Stata code used to produce the output. You should include any relevant plots/summaries you refer to.

Answer the questions in the sections below.

### Q1.1. Initial variable examination:

Examine each variable on its own. Provide an appropriate numerical (descriptive statistics) summary (or summaries) for each variable, as well as an appropriate plot. Present these summaries in a neat and concise way.

**Hint!** A super nifty and easy way to make pretty tables is to use the stata asdoc command. See https://fintechprofessor.com/2018/01/31/asdoc/ for instructions

### Q1.2. Covariate Screening

We know that characteristics other than sex may also be related to salary and so we will want to screen for potential confounders (we’ll talk more about this in coming weeks). But, we need to examine bivariate relationships between salary and: degree, rank, and years from degree. We also need to test for bivariate associations between sex and degree, rank, and years from degree.

A formal confounder is a variable that is associated with both the independent (sex) and dependent (salary) variable.

Present the results of these tests (**hint: there should be six tests**) and indicate which variables are potential confounders.

# Q3. Unadjusted hypothesis test

## Test the hypothesis that the mean salary for the sexes is equal versus an alternative hypothesis that they are not equal (unadjusted—that is, no additional covariates besides the primary predictor).

### Q3.1. State your Research hypothesis

### Q3.2. Identify the assumptions, and are they met?

### Q3.3. Problematic assumptions aside, why might such a simple test be inadequate to test for differences between men and women?

### Q3.4. Carry out the unadjusted test of your hypothesis (assume that the assumption is met). Paste the result and interpret.

### **Extra Credit:** Assume that the key assumption above was violated. What alternative test could you use to test for differences between sexes? Paste the results of that test here

## Testing the hypothesis another way

### Q3.5. Fit an unadjusted regression model to test if the mean salary for the sexes differs. What do you conclude here? How do you interpret the relationship between salary and gender? Any caveats?