

Methods Subfield Questions

Pol211, UC Davis

Hanno Hilbig

Problem Set 1

In this problem set, we will analyze the relationship between various demographic traits and pro-feminist voting behavior among circuit court judges. In a paper, Adam N. Glynn and Maya Sen argue that having a female child causes circuit court judges to make more pro-feminist decisions. The paper can be found at:

Glynn, Adam N., and Maya Sen. (2015). "Identifying Judicial Empathy: Does Having Daughters Cause Judges to Rule for Women's Issues?." *American Journal of Political Science* Vol. 59, No. 1, pp. 37–54.

The dataset `dbj.csv` contains the following variables about individual judges:

Name	Description
<code>name</code>	The judge's name
<code>child</code>	The number of children each judge has.
<code>circuit.1</code>	Which federal circuit the judge serves in.
<code>girls</code>	The number of female children the judge has.
<code>progressive.vote</code>	The proportion of the judge's votes on women's issues which were decided in a pro-feminist direction.
<code>race</code>	The judge's race (1 = white, 2 = African-American, 3 = Hispanic, 4 = Asian-American).
<code>religion</code>	The judge's religion (1 = Unitarian, 2 = Episcopalian, 3 = Baptist, 4 = Catholic, 5 = Jewish, 7 = Presbyterian, 8 = Protestant, 9 = Congregationalist, 10 = Methodist, 11 = Church of Christ, 16 = Baha'i, 17 = Mormon, 21 = Anglican, 24 = Lutheran, 99 = unknown).
<code>republican</code>	Takes a value of 1 if the judge was appointed by a Republican president, 0 otherwise. Used as a proxy for the judge's party.
<code>sons</code>	The number of male children the judge has.
<code>woman</code>	Takes a value of 1 if the judge is a woman, 0 otherwise.
<code>X</code>	Indicator for the observation number.
<code>yearb</code>	The year the judge was born.

Note: if you are asked to provide an interpretation or an explanation, 1-3 sentences are sufficient. There is no need to write more than that. Short, succinct answer are preferred.

Question 5 (3 points)

Load the `dbj.csv` file. Our outcome in this exercise will be the proportion of pro-feminist decisions, `progressive.vote`. What is the difference in the proportion of pro-feminist decisions between judges who have at least one daughter and those who do not have any? Compute this difference in two ways; (1) using any judge who has children, (2) separately for judges that one, two, or three children. For (2), you should

end with three estimates: one for the judges with one child, one for the judges with two children, and one for the judges with three children (HINT: you will subset the data quite a few times to achieve this).

Problem Set 2

Question 5 (3 points)

Creating descriptive statistics is a key part of any research project. We will work with data compiled by Raj Chetty and coauthors. The paper is here: <https://www.nber.org/papers/w23618>. Alternatively, you can also read a summary report here: https://opportunityinsights.org/wp-content/uploads/2018/03/coll_mrc_summary.pdf). The purpose of this data is measure the role of colleges in facilitating upward (economic) mobility.

1. Using a for loop, calculate the mean and standard deviation of the following variables in the Chetty data: `sat_avg_2013`, `par_median`, `par_top1pc`. The result should be a data frame with 3 columns: `variable`, `mean` and `sd`, and three rows, one for each variable.
2. In R, an alternative to for loops are the `apply` family of functions. Repeat subquestion 1 using the `apply` function of your choice (eg `apply`, `sapply`, `lapply`). The output should be the same as for subquestion 1.

The standard deviation measures dispersion of the data. However, we can also create another measure of the dispersion of the sample mean, called the **standard error of the mean**. It is defined as follows: $s_{\bar{x}} = \frac{s_x}{\sqrt{N}}$, where s_x is the standard deviation of a variable x , and N is the number of observations.

3. Write a short function that calculates the SE of the mean, and takes a variable `x` as the argument. Then, calculate the mean and the SE of the mean for `par_median` separately for CA, NY, TX and FL. Finally, create a plot that has the name of each state on the x-axis. It should then show the mean of `par_median` as a point on the y-axis, and also show error bars that visualize the following interval: $[\bar{x} - 1.96s_{\bar{x}}, \bar{x} + 1.96s_{\bar{x}}]$
- I recommend doing this in `ggplot` using the functions `geom_point` and `geom_errorbar`

Problem Set 3

Question 5 (3 points)

5.1

In lecture, we said the mean $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the best “guess” of x_i in terms of minimizing the mean squared error. In statistics, it is common to define some measure of how “good” a guess or an estimate is, and then to show that there is some specific estimate that is “best” in terms of this measure. Recall that the mean squared error is defined as

$$MSE = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x})^2$$

where \hat{x} is our guess of x_i .¹

- Show that the mean \bar{x} is the best guess in terms of minimizing the mean squared error given above.
- To show this, you will use calculus. Recall that we can use derivatives to find minima / maxima of a function. Commonly, we proceed as follows: (1) we take the derivative of a function with respect to the variable we want to optimize over, (2) we set the derivative equal to zero, and (3) we solve for the variable we want to optimize over. In this case, we want to optimize over \hat{x} , so we will take the derivative of the mean squared error with respect to \hat{x} , set it equal to zero, and solve for \hat{x} . The goal is then to show that $\hat{x} = \bar{x}$ is the solution to this problem.
- **Note 1:** Usually, we would also have to assess whether we found a minimum or maximum by looking at the second derivative. However, you can just assume that you found a minimum here – looking at the second derivative is not necessary.
- **Note 2:** You can use the fact that $\sum_{i=1}^n a = na$, since a is just a constant.

5.2

In linear regressions, goodness of fit is usually measured using the R^2 statistic. This statistic is defined as follows:

$$R^2 = 1 - \frac{SSR}{SST}$$

where the sum of squared residuals $SSR = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ and the total sum of squares $SST = \sum_{i=1}^n (y_i - \bar{y})^2$. Here, \hat{y}_i are the predicted values of y_i from the regression, and y_i are the observed values of Y .

In R, write a function that calculates R^2 for a given regression, based on the definition given above. The function should have two arguments: (1) the observed values y_i and (2) an `lm` object that contains the regression results of a regression of Y on X . The function should return the R^2 statistic.

```
x <- c(2, 5, 3, 6, 7, 8, 3, 6, 7, 9)
y <- c(9, 6, 8, 4, 7, 5, 4, 3, 3, 2)

mod <- lm(y ~ x)

get_rsqr <- function(y, lm_object) {

  # Your code here

}
```

¹Note that \hat{x} is usually some function of the data, i.e. $\hat{x} = f(x_1, x_2, \dots, x_n)$. Quantities like the mean, median or mode are all functions of the data.

5.3

If R^2 is closer to 1, this is usually considered a better fit.

1. First, provide an intuitive explanation why R^2 is a measure of goodness of fit. For this, it is helpful to consider cases where SSR is small relative to SST, and cases where SSR is large relative to SST.
2. Second, please state whether R^2 tells us something about the direction of the relationship between X and Y .
3. Third, assume we are interested in explaining whether citizens vote for Democrats or Republicans. Many different variables are correlated with voting behavior, such as age, income, gender, race, etc. For many of these variables, we can find a “statistically significant”² relationship with voting behavior. However, the R^2 of regressing voting behavior on these variables is usually quite low (e.g. < 0.2), which indicates that a given X only explains a small part of the variation in Y . However, researchers in social science generally do not consider this a problem. Please explain why not.

²We will cover what this means in more detail later, but I am sure you heard this term before.

Problem Set 4

Question 6 (3 points)

6.1

Assume we have a continuous random variable with the following probability density function (PDF):

$$f(x) = \begin{cases} ax & \text{if } 0 \leq x \leq 2 \\ 0 & \text{otherwise} \end{cases}$$

Here, a is a constant (i.e. just a number). One requirement for this to be a proper probability density function is that the integral over all possible values of X is equal to 1. In other words, we need to have:

$$\int_0^2 f(x)dx = 1$$

Note that the integral has to be over all possible values of x , which in this case is $0 \leq x \leq 2$.

- First, identify which value the constant a has to have to make this a proper probability density function.
- Then, calculate the expected value of X .

6.2

Assume you have the following continuous joint PDF for the variables X and Y

$$f(x, y) = \begin{cases} \frac{3}{2}y^2 & \text{if } 0 \leq x \leq 2 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- What are the marginal probability density functions of X and Y ?
- Given what you found, are the two variables independent?

Hint: When you integrate over one of the two variables in a joint distribution, then the limits of integration should be the range of values that the variable can take on. For example, if you want to integrate over X , then the limits of integration should be $0 \leq x \leq 2$.

6.3

Now, let's look at the following joint distribution. Note that c is just a constant.

$$f(x, y) = \begin{cases} c(x + y^2) & \text{if } 0 \leq x \leq 1 \text{ and } 0 \leq y \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- Find the conditional distribution of X given Y , i.e. $f_{X|Y}(x|y)$.
- Then, calculate the following quantity: $Pr[X < \frac{1}{2} | Y = \frac{1}{2}]$.

Note: Unlike in 6.1, it is not necessary to find the value of c to answer this question. You can just treat c as a constant throughout the calculations. If you do everything correctly, you will see that c eventually disappears from the final expression.

Hint: To calculate the second quantity, you can just plug $Y = 1/2$ into the conditional distribution you found, and use integration to find the desired quantity.

Problem Set 5

Question 5 (3 points)

5.1 (1 point)

Assume that we have a simple linear regression of the form $E(Y|X) = \alpha + \beta X$. The linear regression gives us an estimate of the expected value of Y given X .

Note that the coefficients α and β are chosen to minimize the following expression:

$$E[(Y - \alpha - \beta X)^2]$$

In other words, the coefficients are chosen to minimize the expectation of the squared difference between the expectation of Y given X and the actual value of Y . We can use the last expression to derive the following expression for β and α :

$$\begin{aligned}\beta &= \frac{Cov(X, Y)}{Var(X)} \\ \alpha &= E(Y) - \beta E(X)\end{aligned}$$

Show that the expressions of β and α given above minimize the expression $E[(Y - \alpha - \beta X)^2]$. To do this, you can take the derivatives of the expression in (1) with respect to α and β , and set the derivatives to 0. You can then solve for α and β .³

Note: For this question, you can assume the following:

$$\frac{\partial E[f(x, y)]}{\partial(x)} = E \left[\frac{\partial f(x, y)}{\partial x} \right]$$

Where $\frac{\partial f(x, y)}{\partial x}$ is the partial derivative of some function $f(x, y)$ with respect to x .

Hint 1: For some constant a and RVs X and Y :

- $E(aX) = aE(X)$
- $E(a) = a$
- $E(Y + X) = E(Y) + E(X)$

Hint 2: $Var(X) = E(X^2) - E(X)^2$

Hint 3: $Cov(X) = E(XY) - E(X)E(Y)$

5.2 (0.5 points)

Next, recall that the definition of the correlation coefficient is:

$$\rho(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}$$

- Mathematically, what is the relationship between the correlation coefficient and the slope of the bivariate linear regression?
- When are the two the same?
- Can you provide a brief intuition / explanation for the condition under which the two are the same?

³As in a previous problem set, we will ignore the second derivative, i.e. we will assume we found a minimum.

5.3 (0.5 points)

Consider the following PDF:

$$f(x) = \begin{cases} \frac{3}{2}x^2 & \text{if } -1 \leq x \leq 1 \\ 0 & \text{otherwise} \end{cases}$$

- a. What is $E(X)$? Please do not use the integrals here, but rather come up with an intuitive explanation based on the shape of the distribution.
- b. What is the expected value of the new random variable Y , which is defined as the absolute value of X , i.e. $Y = |X|$?

5.4 (0.5 points)

Suppose that a point is chosen at random on a stick of length 1 and that the stick is broken into two pieces at that point. Find the expected value of the length of the longer piece.

5.5 (0.5 points)

Use a simulation to check your answer to question 5.4.

- First, simulate 1,000 draws to verify whether your answer from the previous question is correct.
- Then, write code that simulates \bar{Y}_n , which is the mean of Y as a function of the number of draws n from the uniform RV X . Here, $n \in \{10, 20, 30, \dots, 500\}$. What do you notice about the relation between \bar{Y}_n and n , in particular with respect to $E(Y)$, which you derived in question 5.4?

Problem Set 6

Question 7 (3 points)

7.1 (1 point)

An important result in statistics is the weak law of large numbers (WLLN), which we will now prove. It states the following:

$$\lim_{n \rightarrow \infty} P(|\bar{X}_N - \mu| > \epsilon) = 0$$

where \bar{X}_N is the sample mean of a sample with N iid draws from an RV X with population mean μ . The WLLN says that the probability that the sample mean is more than ϵ away from the population mean goes to zero as the sample size goes to infinity. This is called convergence in probability.

a. As a first step, prove the following, where \bar{X}_N is the sample mean of a sample with N draws from an RV X with population mean μ and finite variance σ^2 :

$$\lim_{n \rightarrow \infty} E[(\bar{X}_N - \mu)^2] = 0$$

Hint 1: You can use the fact that $E(\bar{X}_N) = \mu$

Hint 2: The proof is short.

b. An important inequality in probability theory is Chebyshev's inequality. It states the following:

$$P(|Y| \geq \epsilon) \leq \frac{E(Y^2)}{\epsilon^2}$$

Here, Y is a random variable and ϵ is a positive number.

Use Chebyshev's inequality and the result from **a.** to prove the weak law of large numbers, which is stated above.

Finally, briefly explain why the variance of the original RV has to be finite for the weak law of large numbers to hold.

7.2 (1 point)

Assume you have two independent uniform RVs distributed as follows:

$$Y \sim \text{Unif}(0, 1)$$

$$X \sim \text{Unif}(1, 2)$$

Assume we draw a sample from each, where the sample size of each sample is N . We are now interested in $\bar{X} - \bar{Y}$.

a. If you want to ensure that, in 95% of the cases, your estimate of $\bar{X} - \bar{Y}$ is within 0.1 of the population difference (i.e. $E(X) - E(Y)$), what is the minimum required sample size N to achieve this?

b. Verify your answer using a simulation in R.

7.3 (1 point)

For a linear regression model with observed values x_i and y_i , the least squares estimator for the slope β is given by:

$$\hat{\beta} = \frac{\sum_{i=1}^N (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^N (x_i - \bar{x})^2}$$

where \bar{x} and \bar{y} are the sample means of x_i and y_i , respectively. Let's define $S_X^2 = \sum_{i=1}^N (x_i - \bar{x})^2$. In addition, the variance of the RV Y is σ^2 , and the draws y_i are iid. Note that we can also write the above equation as:

$$\hat{\beta} = \frac{\sum_{i=1}^N (x_i - \bar{x})y_i}{S_X^2}$$

Find the variance of the least squares estimator $\hat{\beta}$. Note that for the purposes of this problem, we can treat the values x_i as constants, i.e. not as random variables.⁴

⁴This is a convention that is often used in statistics – we consider β to be a function of the RV Y , but not of the RV X .

Problem Set 7

Question 6 (3 point2)

Q6 builds upon the context established in Q5. The framework set up in Q5 is as follows:

In a randomized experiment, researchers want to evaluate an intervention that informs young people about the importance of participation in local politics. The outcome that the researchers are interested in is the amount of hours that the participants spend participating in local politics, which includes activities such as attending city council meetings, working for local political campaigns, and so on.

Individuals are randomly assigned to the intervention (the “treatment group”) or to a control group where they are exposed to information unrelated to politics. After the experiment, the researchers observe that the average number of hours spent participating in local politics is 2.3 hours per week for the treatment group, and 2 hours per week for the control group. The variance of the number of hours spent participating in local politics is 8 for both groups. The sample size for both groups is 400.

For the rest of the problem, let the sample mean for the treatment group be \bar{X}_T , and the sample mean for the control group be \bar{X}_C .

6.1 (1 point)

Assume the same setup as in question 5, with $N_T = N_C = 900$. We now want to calculate the *power* of the test. Power is defined as the probability of rejecting the null hypothesis when the null hypothesis is false.⁵ In particular, we will now assume that the true difference between the population means is 0.3 hours per week, i.e. $\mu_T - \mu_C = \Delta\mu = 0.3$. Further, we choose the significance level to be 0.05.

- a. Under the null hypothesis in question 4, what are the values for which we reject the null, using the significance level given above? Show this region graphically in a plot of the distribution of the sample mean under the assumption that the null hypothesis holds.
- b. Consider scenario outlined above, where $\Delta\mu = 0.3$. What is the distribution of the difference between the sample means under this scenario? Show this distribution graphically. In the same plot, visualize the rejection region you derived in part a..
- c. Under the assumption that the true $\Delta\mu = 0.3$, what is the probability of rejecting the null hypothesis? This is the power of the test. Interpret the quantity you calculated in words.
- d. The previous steps only gave us the power for one specific value of $\Delta\mu$. In practice, we are interested in the power for a range of values of $\Delta\mu$. Use R to plot the power of our test for values of $\Delta\mu$ between -0.5 and 0.5. Interpret the shape of the curve.
- e. The previous question asked you to show how power depends on the true difference in means $\Delta\mu$. Assuming that $\Delta\mu$ is fixed, what are some other ways of increasing the power of the test described in question 4? Name at least two.
- f. We say that we make a type II error if we fail to reject the null when it is false. In the context of this question, what is the probability of making a type II error if $\Delta\mu = 0.3$?

6.2

We will continue with the example from question 5. We assume the alternative hypothesis is correct, i.e. $\Delta\mu = 0.3$. We further assume a significance level of 0.05 and standard error of the difference in means that is $SE(\bar{X}) = 4/\sqrt{N}$. However, our goal is now to determine the size of each group N for a one-tailed test that has a power of 0.9.

⁵Making statements like “the null hypothesis is false” requires us to make some assumptions about the true population parameters.

We now denote $\Delta\mu_0 = 0$ as the difference in means under the null hypothesis, and $\Delta\mu_A = 0.3$ as the difference in means under the alternative hypothesis. For a one-tailed test to have a power of 0.9, the following needs to hold:

$$P\left(\frac{\Delta\bar{X} - \Delta\mu_0}{4/\sqrt{N}} > 1.645\right) = 0.9$$

So we are just saying: the probability of the standardized difference in sample means being greater than the critical value of 1.645 needs to be 0.9. For a one-tailed test with the significance level of 0.05, the critical value is 1.645.

Determine the required sample size N for a power of 0.9 by solving the equation above for N .

Hint 1: This calculation should be done assuming that the alternative hypothesis is true, which implies that $E(\Delta\bar{X}) = \Delta\mu_A = 0.3$.

Hint 2: The inverse of the CDF is called the quantile function. In R, this function is called `qnorm`.

Problem Set 8

Question 5 (3 points)

An alternative way to test hypotheses without having to make assumptions about the distribution of the test statistic is **randomization inference**.

Let Y_i be a binary outcome for unit i , and D_i be a binary treatment assignment for unit i . As before, the potential outcomes are Y_{1i} and Y_{0i} . We now test a sharp null hypothesis of $H_0 : Y_{1i} = Y_{0i}$ for all i .

For our test statistic, we will use the absolute difference in means estimator, which is defined as follows:

$$T = \left| \frac{1}{N_1} \sum_{i=1}^N D_i Y_i - \frac{1}{N_0} \sum_{i=1}^N Y_i (1 - D_i) Y_i \right|$$

where N is the total number of units, N_1 is the number of units in the treatment group, and N_0 is the number of units in the control group.

Our data looks like this:

Case	Y_i	D_i
1	1	1
2	1	1
3	1	1
4	0	0
5	1	0
6	0	0

In R, we can calculate the test statistic as follows:

```
# data

Y = c(1, 1, 1, 0, 1, 0)
D = c(1, 1, 1, 0, 0, 0)

# number of units

get_T <- function(Y, D) {

  term1 <- mean(Y[D == 1])
  term2 <- mean(Y[D == 0])

  ## For some possible treatment vectors, the mean of Y[D == 1] or Y[D == 0] may be NaN
  ## In those cases, just set the mean to zero
  ## This is fine since the other term will simply be the mean of Y across all units

  term1 <- ifelse(is.nan(term1), 0, term1)
  term2 <- ifelse(is.nan(term2), 0, term2)

  abs(term1 - term2)

}

get_T(Y, D)
```

[1] 0.6666667

The observed T is equal to $2/3$.

a. Assume that the null hypothesis holds, but the treatment assignment D_i is different from the one observed above, i.e. some cases receive the treatment that currently do not, and some cases do not receive the treatment that currently do. Under a different treatment assignment, what would the observed values of Y_i be?

b. Assume each unit is able to receive the treatment or not. Let D be a vector that contains the treatment assignment for each unit, which may be different from the one we actually observe. In the observed data, $D = (1, 1, 1, 0, 0, 0)$. How many different vectors D are there in total?

c. Assuming that the null holds, is it possible to calculate T for a different treatment vector D with the observed data given above (i.e. a vector different from the one observed above), or does the fundamental problem of causal inference prevent this?

d. In R, create all possible permutations of the vector D . For each, calculate the test statistic T . This gives you the distribution of T if the null hypothesis is true.

e. We can now calculate a p-value, which tells us how “typical” our observed value of T is under the null hypothesis. Let \tilde{T}_k be test statistic that is derived from the k^{th} permutation of D . The p-value is then defined as follows:

$$p = \frac{1}{K} \sum_{k=1}^K 1(\tilde{T}_k \geq T)$$

where K is the total number of permutations, and $1(\cdot)$ is the indicator function.

Calculate the p-value under the null. Using the typical threshold of $\alpha = 0.05$, can you reject the null hypothesis?

f. What is the advantage of randomization inference over the CLT-based approach we have used so far, particularly in cases like the one above?

g. Assume a similar scenario with $N = 4$. In this case, is it possible to reject the null at the 95% confidence level? Why or why not?

h. Unlike in the example above, assume that N is large. Even with a modest sample size of 100, the number of possible permutations of D is $> 10^{30}$. This makes it infeasible to calculate the p-value as we did above. How can we still conduct randomization inference in this case?