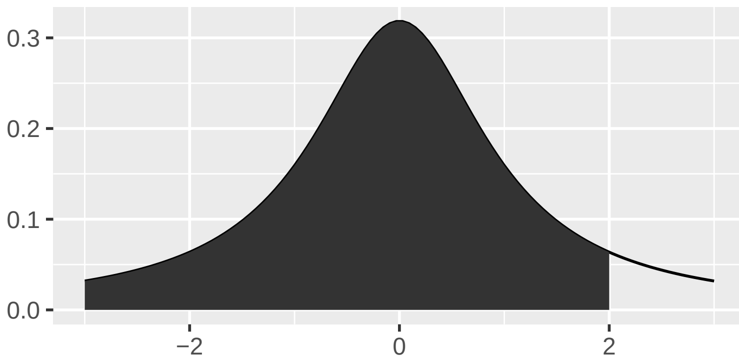
Exam 1 Practice Solutions

Econ B2000, MA Econometrics

Shay Culpepper, CCNY

Fall 2018



Using the Normal and Student's T to find p-values

What will be helpful in this section: `pnorm`, `qnorm`, `pt`, `qt`, and a normal distribution / students t distribution graph to visualize. 2 ways to do it. 1) You can calculate the z score. $z = \frac{\bar{x}-\mu}{\sigma}$, or specify mean and sd in the function itself. pnorm and pt default to the lower tail.

Please answer the following. You may find it useful to make a sketch

Example set 1

a. `pnorm(20.84, mean = 11, sd = 8.2, lower.tail = FALSE) ## 0.1150697`
b. `pnorm(-6.6, mean = -5, sd = 4) ## 0.3445783`
c. `pnorm(3.45, mean = 12, sd = 4.5) ## 0.02871656`
d. `pnorm(-28.82, mean = -14, sd = 7.8, lower.tail = FALSE) ## 0.9712834`
e. `pnorm(-3.7, mean = 8, sd = 9) ## 0.09680048`
f. `2 * pnorm(-14.96, mean = -10, sd = 6.8) ## 0.4657498`
g. `2 * pnorm(-6 - 12.27 + -6, mean = -6, sd = 4.2) ## 1.361375e-05`
h. `2 * pnorm(10 - 2.8, mean = 10, sd = 6.4) ## 0.6617488`
i. `c( qnorm(0.118 / 2, mean = -6, sd = 3.5), -qnorm(0.118 / 2, mean = -6, sd = 3.5)) ## -11.47128  11.47128`
j. `c( qnorm(0.024 / 2, mean = -4, sd = 0.1), -qnorm(0.024 / 2, mean = -4, sd = 0.1)) ## -4.225713  4.225713`
k. `2 * pt(-4.32 / 2.7, df = 12) ## 0.1355805`
l. `2 * pt(-19.11 / 9.1, df = 40) ## 0.04208202`
m. `2 * pt(-21.16 / 9, df = 29) ## 0.02572611`

Example set 7

a. `pnorm(2.1)  ## D. 0.9821`
b. `pnorm(-0.6) ## A. 0.2743`
c. `pnorm(0.3)  ## C. 0.6179`
d. `pnorm(0.9, lower.tail = FALSE)  ## A. 0.1841`

e. `pnorm(-0.4, lower.tail = FALSE) ## D. 0.6554`
f. `2 * pnorm(-1.8)  ## D. 0.0719`
g. `2 * pnorm(-0.5)  ## D. 0.6171`
h. `2 * pnorm(-2.4)  ## C. 0.0164`
i. `qnorm(0.324 / 2) ## C. +-0.986`
j. `qnorm(0.390 / 2) ## A. +-0.8596`
k. `qnorm(0.218 / 2) ## C. +-1.2319`

Example set 9

a. `2 * pnorm(-1.9) ## 0.05743312`
b. `2 * pnorm(-1.5) ## 0.1336144`
c. `2 * pnorm(-1.2) ## 0.2301393`

Example set 10

a. `2 * pnorm(-3,  mean = -1, sd = 1.5) ## 0.1824224`
b. `2 * pnorm(-45, mean = 50, sd = 30)  ## 0.00154197`
c. `## 1`

Example set 11

a. `2 * pnorm(1.75, lower.tail = FALSE) ## 0.08011831`
b. `2 * pnorm(2,    lower.tail = FALSE) ## 0.04550026`
c. `2 * pnorm(1.3,  lower.tail = FALSE) ## 0.193601`
d. `2 * pnorm(2.1,  lower.tail = FALSE) ## 0.03572884`
e. `c(qnorm(.1),   qnorm(.9))    ## -1.281552  1.281552`
f. `c(qnorm(.05),  qnorm(.95))   ## -1.644854  1.644854`
g. `c(qnorm(.025), qnorm(.975))  ## -1.959964  1.959964`

Example set 12

a. `pnorm(0) - pnorm(-1.75) ## 0.4599408`
b. `pnorm(1.75) - pnorm(0) ## 0.4599408`
c. `2 * pnorm(-1.75) ## 0.08011831`
d. `c(qnorm(.05),  qnorm(.95))   ## -1.644854  1.644854` & `c(qnorm(.025), qnorm(.975))  ## -1.959964  1.959964`

Example set 13

a. `pnorm(7,  mean = 3, sd = 4) - pnorm(3, mean = 3, sd = 4) ## 0.3413447`
b. `pnorm(11, mean = 3, sd = 4) - pnorm(7, mean = 3, sd = 4) ## 0.1359051`
c. `2 * pnorm(-7,  mean = 3, sd = 4)  ## 0.01241933`

Statistics from given numbers (no datasets in R required)

1. *** Confidence Intervals, Hypothesis Tests

a.

```r
phat <- 0.54
n <- 200
critval <- -qnorm(.05)
E <- critval * sqrt( phat * (1 - phat) / n )
```

```r
c(phat - E, phat + E)
```

```
## [1] 0.482032 0.597968
```

b.
```r
phat <- 0.54
n <- 300
critval <- -qnorm(.05)
E <- critval * sqrt( phat * (1 - phat) / n )

c(phat - E, phat + E)
```

```
## [1] 0.4926694 0.5873306
```

c.
```r
phat1 <- 0.54
n1 <- 200

phat2 <- 0.51
n2 <- 200

pbar <- (phat1 * n1 + phat2 * n2)/ (n1 + n2)
critval <- -qnorm(.05)
test.stat <- (phat1 - phat2) / sqrt( pbar * (1 - pbar) * (1/n1 + 1/n2) )

c(critval, test.stat)
```

```
## [1] 1.6448536 0.6007514
```

```r
abs(test.stat) > abs(critval)
```

```
## [1] FALSE
```

d. Candidate X must win 2 particular states in order to win the election; the forecast says she has a 60% chance of winning each state individually. Your friend, a wannabe statistician, explains that a 0.6 chance of winning one state and a 0.6 chance of winning the other means only a 0.6*0.6= 0.36 chance of winning both - so the "favorite" is actually not the favorite! Explain why your friend is wrong. ** I have questions on this. Does 60% mean that the polls are at 60%, or does it mean that the likelihood that the support for X is >= 0.51 is 60% based on their polling results?

2. *** Confidence Intervals, Minimum n (20 points)

Suppose that a particular medical treatment already improves patient outcomes by 20 (don't worry about the units for now) and it is established that the standard deviation for the population is 8. There is an improved treatment that is expected to deliver a further 10% improvement.

a. If there were 10 patients in the trial, what would be the t-statistic, p-value, and confidence interval - assuming the new treatment works as expected? Carefully explain the null hypothesis.

```r
x.bar <- 30
n <- 10
critval <- -qnorm(0.05)
E <- critval * (8 / sqrt(10))
test.stat <- 10 / (8 / sqrt(10))

pnorm(-test.stat)
```

```
## [1] 3.86134e-05
```

The null hypothesis is that the improvement from this treatment improves patient outcomes by 20.

$$H_0 : \mu \leq 20$$
$$H_a : \mu > 20$$

b. If there were 30 patients, what would be the t-stat, p-value, and confidence interval (again assuming the treatment works as expected)?

```
x.bar <- 30
n <- 10
critval <- qnorm(0.05)
E <- critval * (8 / sqrt(30))
test.stat <- 10 / (8 / sqrt(30))

pnorm(-test.stat)
```

```
## [1] 3.783087e-12
```

c. If the company wants a p-value of 5% or lower, how many patients should they plan to have in the trial? *What is the desired margin of error??*

```
(-pnorm(0.025) * 8) ** 2
```

```
## [1] 16.64461
```

3. Hypothesis Tests, Conditionals (20 points)

a. Testing whether the fraction of immigrants of people making less than 15 and hour vs the fraction of immigrants of people making more than 15 an hour

```
n1 <- 14235 + 3113 + 3113 + 1824
x1 <- 3113 + 1824
phat1 <- x1 / n1

n2 <- 33150 + 662 + 5296 + 567
x2 <- (5296 + 567)
phat2 <- x2 / n2

pbar <- (x1 + x2) / (n1 + n2)
t.stat <- (phat1 - phat2) / sqrt( pbar * (1 - pbar) * (1/n1  + 1/n2))
critval <- pnorm(0.025)
p.val <- pnorm(-t.stat)
E <- critval * sqrt( phat1 * (1 - phat1) / n1 + phat2 * (1 - phat2 ) / n2)
point.est <- phat1 - phat2
```

$$z = 23.2265503$$
$$p\text{-value} = 1.2278041 \times 10^{-119}$$
$$E = 0.0016847$$
$$0.0720788 \leq p_1 - p_2 \leq 0.0754482$$

a. Of immigrants, the fraction who are making \$15/hr vs who are making more than \$15/hr. In this case we'll test the proportion of immigrants making less than \$15/hr against the null hypothesis being that the proportion equals .50.

```r
n <- 3113 + 1824 + 5296 + 567
phat <- (3113 + 1824) / n
p <- 0.5

critval <- -pnorm(0.005)
point.est <- phat
se.phat <- sqrt( phat * (1 - phat) / n )
se.p <- sqrt( p * (1 - p) / n )
E <- critval * se.phat
t.stat <- point.est * se.p
p.val <- pnorm(abs(t.stat)) * 2
```

$$
\begin{aligned}
z &= 0.0021994 \\
p\text{−value} &= 1.0017548 \\
E &= -0.0024063 \\
0.459536 \leq p &\leq 0.4547233
\end{aligned}
$$

b.
```r
n1 <- 14235 + 3113 + 1062 + 1824
x1 <- 1062 + 1824
phat1 <- x1 / n1

n2 <- 33150 + 662 + 5296 + 567
x2 <- (662 + 567)
phat2 <- x2 / n2

pbar <- (x1 + x2) / (n1 + n2)
t.stat <- (phat1 - phat2) / sqrt( pbar * (1 - pbar) * (1/n1  + 1/n2))
critval <- pnorm(0.025)
p.val <- pnorm(-t.stat)
E <- critval * sqrt( phat1 * (1 - phat1) / n1 + phat2 * (1 - phat2 ) / n2)
point.est <- phat1 - phat2
```

$$
\begin{aligned}
z &= 51.1026065 \\
p\text{−value} &= 0 \\
E &= 0.0013299 \\
0.1103247 \leq p_1 - p_2 &\leq 0.1129844
\end{aligned}
$$

b.
```r
n <- 1062 + 1824 + 662 + 567
phat <- (1062 + 1824) / n
p <- 0.5

critval <- -pnorm(0.005)
point.est <- phat
se.phat <- sqrt( phat * (1 - phat) / n )
se.p <- sqrt( p * (1 - p) / n )
E <- critval * se.phat
t.stat <- point.est * se.p
p.val <- pnorm(abs(t.stat)) * 2
```

$$z = 0.0054665$$
$$p\text{−value} = 1.0043616$$
$$E = -0.0035815$$
$$0.7049181 \leq p \leq 0.697755$$

c. `1824 / (14235 + 3113 + 1062 + 1824) ## 0.0901453`
d. `1824 / (3113 + 1824) ## 0.3694551`
e. `567 / (5296 + 567) ## 0.09670817`

4. Confidence Intervals, Hypothesis Tests (20 points)

a.
```
n1  <-  325
sd1 <- .1513
x1  <- -.0498

n2  <- 162
sd2 <- .1836
x2  <- .0815


critval <- pt(0.025, df = min(n1, n2) - 1)
t.stat <- (x1 - x2) / sqrt( (sd1**2)/n1  + sd2**2/n2)
p.val <- pnorm(t.stat)
E <- critval * sqrt( (sd1**2)/n1  + sd2**2/n2)
point.est <- x1 - x2
```

$$t = -7.867553$$
$$p - \text{value} = 1.8082258 \times 10^{-15}$$
$$E = 0.0085106$$
$$-0.1398106 \leq \mu_1 - \mu_2 \leq -0.1227894$$

b.
```
n1  <-  112
sd1 <- .1431
x1  <- -.0349

n2  <- 75
sd2 <- .1840
x2  <- .0667


critval <- pt(0.025, df = min(n1, n2) - 1)
t.stat <- (x1 - x2) / sqrt( (sd1**2)/n1  + sd2**2/n2)
p.val <- pnorm(t.stat)
E <- critval * sqrt( (sd1**2)/n1  + sd2**2/n2)
point.est <- x1 - x2
```

$$t = -4.0342589$$
$$p - \text{value} = 2.738745 \times 10^{-5}$$
$$E = 0.0128425$$
$$-0.1144425 \leq \mu_1 - \mu_2 \leq -0.0887575$$

c. With the R-code below, can you find other relationships? Do these differences from above seem reasonable?

```r
library(quantmod)
getSymbols(c('INDPRO','UNRATE'),src='FRED')
```

```
## [1] "INDPRO" "UNRATE"
```

```r
ip_1 <- INDPRO["1965::"]
ur_1 <- UNRATE["1965::"]
d_ip <- na.trim(ip_1 - lag(ip_1))
d_ur <- na.trim(ur_1 - lag(ur_1))
```

6. Correlations, Hypothesis Tests (20 points)

A recent research paper, looking at how much attractiveness and personal grooming affects wages, used data from The National Longitudinal Study of Adolescent Health in 2001-2.

a.
```r
phat1 <- 0.388
n1 <- 6074 * 0.484

phat2 <- 0.506
n2 <- 6074 * 0.506

point.est <- phat1 - phat2
pbar <- (phat1 * n1 + phat2 * n2)/ (n1 + n2)
critval <- -qnorm(.05)
test.stat <- (phat1 - phat2) / sqrt( pbar * (1 - pbar) * (1/n1 + 1/n2) )
E <- critval * sqrt( phat1 * (1 - phat1) ) / n1 + phat2 * (1 - phat2) ) / n2 )

c(critval, test.stat)
## 1.6448536 0.6007514

abs(test.stat) > abs(critval)
## False
```

$$z = -4.0342589$$
$$p{-}\text{value} = 2.738745 \times 10^{-5}$$
$$E = 0.0128425$$
$$-0.1144425 \leq p_1 - p_2 \leq -0.0887575$$

or very well groomed; 50.6% of the females were rated that way. Is this a statistically significant difference?

b. The study considers interrelations between physical attractiveness and grooming. People were ranked on a 4-point scale (where 1 is below average, 2 is average, 3 is above average, and 4 is very much above average) for each attribute. The full details are:

Physically

	4 Very Attractive	3 Attractive	2 Average	1 Less Attractive
4 Very well groomed	297	199	57	30
3 Well groomed	290	1169	607	54
2 Average grooming	75	788	2013	167
1 Less than average grooming	1	25	164	138

c. Conditional on a person being ranked physically 3 or 4 in attractiveness (above average), what is the chance that they are above average (3 or 4) in grooming as well. Conditional on being above average physically, what is the chance that they are average or below average (1 or 2) in grooming? Are these statistically significantly different?

```
### Chance of above average grooming conditional on above average attractiveness
x1 <- 297 + 290 + 199 + 1169
n <- x1 + 1 + 75 + 788 + 25
phat1 <- x1 / n

### Chance of below average grooming conditional on above average attractiveness
x2 <- 1 + 75 + 788 + 25
phat1 <- x2 / n

pbar = 1
```

Personality

The study also considers the attractiveness of someone's personality (charisma), with the same 4-point scale. These data are:

	4 Very Attractive	3 Attractive	2 Average	1 Less Attractive
4 Very well groomed	326	171	60	26
3 Well groomed	416	1186	467	51
2 Average grooming	212	966	1729	136
1 Less than average grooming	11	49	184	84

d. Conditional on having an above-average personality, what is the chance that someone has above-average grooming? Conditional on having an above-average personality, what is the chance that their grooming is at or below average? Is there a statistically significant difference?

e. Comment on the study. If overall attractiveness is a combination of these 3 factors, is there evidence that they are gross substitutes or complements in production?

PK Robins, JF Homer, MT French (2011). "Beauty and the Labor Market: Accounting for the Additional Effects of Personality and Grooming,"" Labour, 25(2), pp 228-251.

Regression Analysis from given data (no datasets in R required)

1. Fill in p-values

To investigate an hypothesis proposed by a student, I got data, for 102 of the world's major countries, on the fraction of the population who are religious as well as the income per capita and the enrollment rate of boys and girls in primary school. The hypothesis to be investigated is whether more religious societies tend to hold back women. I ran two separate models: Model 1 uses girls enrollment rate as the dependent; Model 2 uses the ratio of girls to boys enrollment rates as the dependent. The results are below (standard errors in italics and parentheses below each coefficient):

Find the t-stat by dividing the coefficient by the standard error. Find the p-value using the first coefficient in model 1 as an example `2 * pt(-137/18, df=101)`

	Model 1	t-stat	p-value
Intercept	137 (18)	*7.61*	$1.4731845 \times 10^{-11}$
Religiosity	-0.585 (0.189)	-3.095	0.0025448
GDP per capita	0.00056 (0.00015)	3.73	3.1273573×10^{-4}

	Model 2	t-stat	p-value
Intercept	1.12 (0.09)	12.4444	$2.0789262 \times 10^{-22}$
Religiosity	-0.0018 (0.0009)	-2	0.0240926
GDP per capita	0.0000016 (0.0000007)	2.2857	0.0121805

a. Which coefficient estimates are statistically significant? What are the t-statistics and p-values for each?
b. How would you interpret these results?
c. Critique the regression model. How would you improve it?

Statistics using Datasets (R required)

Regression Analysis using Datasets (R Required)