Final Exam Review — Chapters 1–10, 12

NOTE: For additional review problems, see the review lectures for the midterm exams.

Final Exam

- Time: Tuesday December 14, 3-6pm.
- Place: TBA
- Material: Chapters 1–10, 12 (including all lectures, all homework , all labs)
- It will be a **closed-book exam**.
- Bring your **photo ID**, calculator, ruler, pen, pencil, eraser.

NOTE: The following formulas (plus formulas for midterm exams 1 and 2) will be provided. Tables will be provided if needed.

Formulas.

Hypotheses Testing for Population Means and Proportions

Test for mean μ $[H_0 : \mu = \mu_0]$: test statistic $z = \frac{\bar{x}-\mu_0}{\sigma/\sqrt{n}}$ or $t = \frac{\bar{x}-\mu_0}{s/\sqrt{n}}$.

Test for proportion p $[H_0 : p = p_0]$: test statistic $z = \frac{\widehat{p}-p_0}{\sqrt{p_0(1-p_0)/n}}$.

Correlation coefficient

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right).$$

Simple Linear Regression

Model: $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, where ϵ_i are assumed to be independent and have $N(0, \sigma)$ distribution.

Estimates: slope $\hat{\beta}_1 = r\frac{s_y}{s_x}$, intercept $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, and regression line $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Estimate of σ : $s = \sqrt{s^2} = \sqrt{\text{MSE}}$

CI for slope β_1: $\hat{\beta}_1 \pm t^*\text{se}(\hat{\beta}_1)$ (note: df $= n-2$ for the t^* value)

Test for slope [$H_0 : \beta_1 = 0$] : test statistic $t = \frac{\hat{\beta}_1}{\text{se}(\hat{\beta}_1)}$ (note: df $= n-2$)

1. A service station has both self-service and full-service islands. On each island, there is a single regular unleaded pump with two hoses. Let X denote the number of hoses being used on the self-service island at a particular time, and let Y denote the number of hoses on the full-service island in use at that time. The joint probability distribution of X and Y appears as follows:

$f(x,y)$		y = 0	1	2
	0	.10	.04	.02
x	1	.08	.20	.06
	2	.06	.14	.30

(a) What is the probability of $X < 2$ and $Y = 1$?

(b) What is the probability of $X < 2$ given $Y = 1$?

(c) What is the marginal probability distribution of X?

(d) What are the means of X and Y?

(e) Find the covariance of X and Y?

(f) Are X and Y independent? Explain.

2. Suppose X and Y are two discrete random variables.

(a) Show that $\text{cov}\,(X, Y) = E(XY) - E(X)E(Y)$.

(b) Show that $\text{cov}\,(X, Y) = 0$ if X and Y are independent.

3. A national survey of restaurant employees found that 75% reported that work stress had a negative impact on their personal lives. A manager of a chain of restaurants finds that 68 of a random sample of 100 of the chain's employees say that work stress has a negative impact on their personal lives.

(a) Is there evidence at 5% level to conclude that the proportion for this chain of restaurants differs from the value given for the national survey?

(b) Give a 95% CI for the proportion of employees who work for this chain restaurants who believe that work stress has a negative impact on their personal lives.

(c) How many employees are needed to have a 95% CI with a width at most 5%?

4. A 95% CI for a population mean is (28, 35).

(a) Can you reject the null hypothesis that $\mu = 34$ at the 5% significance level? Why?

(b) Can you reject the null hypothesis that $\mu = 36$ at the 5% significance level? Why?

5. Degree of Reading Power (DRP) scores for a random sample of 44 third-grade students in a suburban school district have a mean of 35.1 and standard deviation 11.2. Assume DRP scores are approximately normal. The researcher believes that the mean score μ of all third graders in this district is higher than the national mean, which is 32.

```
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
     DRP |      44    35.09091    1.686853    11.18932    31.68905    38.49277
------------------------------------------------------------------------------
Degrees of freedom: 43

                             Ho: mean(DRP) = 32

     Ha: mean < 32              Ha: mean != 32              Ha: mean > 32
       t =   1.8324               t =   1.8324               t =   1.8324
   P < t =   0.9631         P > |t| =   0.0738           P > t =   0.0369
```

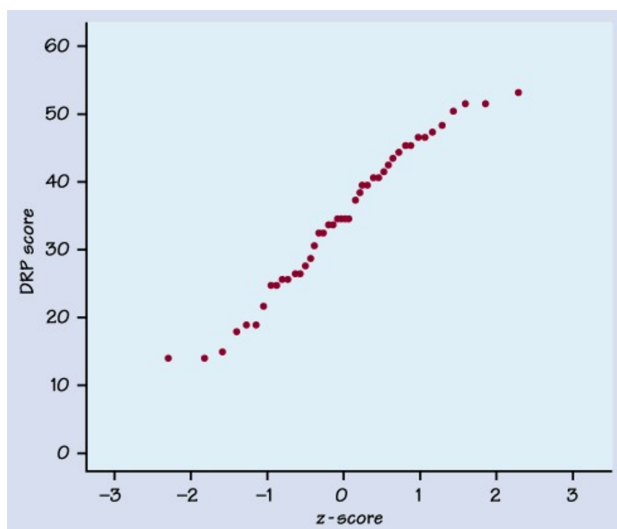
(a) State the appropriate hypotheses to test this suspicion.

(b) What are observed test statistic, degrees of freedom, and p-value?

(c) Give the decision and conclusion at the 5% significance level.

(d) Give the 95% CI for the mean DRP score μ of all third graders in this district.

(e) We assume DRP scores are approximately normal. Comment on the validity of this assumption based on the normal quantile plot for the DRP scores.



6. We wish to assess the effectiveness of a diet program. The weights before and after treatment were recorded for a random sample of 6 people on the program. The data are provided below.

```
Before 175 188 192 200 186 162
After  172 190 186 199 180 163
diff    -3   2  -6  -1  -6   1
```

Is there evidence (at the 10% significance level) to conclude that the program is effective at reducing weight overall?

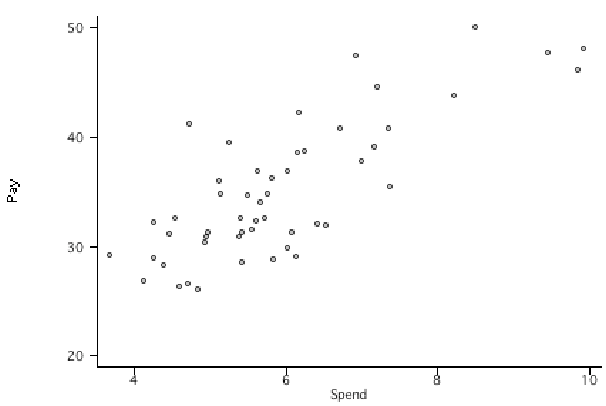
(a) Assume the differences are normally distributed. Perform an appropriate test.

(b) Do not assume the differences are normally distributed. Perform an appropriate test.

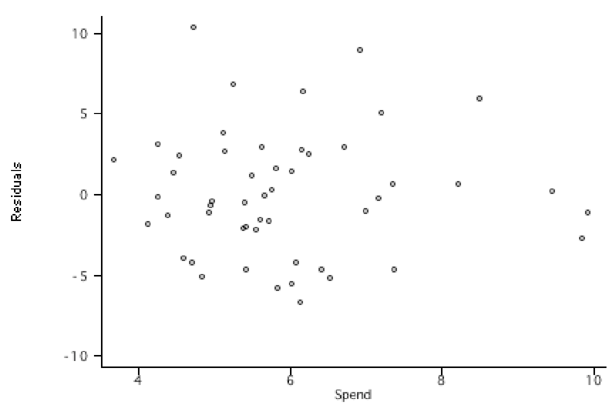
7. Amount of money (thousands of dollars) spent on education by states. Two variables are Pay (average salary paid to teachers) and Spend (expenditures per pupil). We are interested in how the teachers' salary is related to the spending per pupil. Data analysis is done with Stata as follows.

```
. summarize

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         Pay |        51    35.03529    6.268551         26         50
       Spend |        51    5.942549    1.409631       3.67       9.93

. corr
(obs=51)

             |      Pay    Spend
-------------+------------------
         Pay |   1.0000
       Spend |   0.7931   1.0000
```

Scatterplot



Residual plot



(a) Interpret the scatterplot. What do you see?

(b) What are the response variable (y) and explanatory variable (x)?

(c) What is the regression line?

(d) What is the expected teachers' salary if the spending per pupil is \$6,000?

(e) What proportion of the variation in teachers' salary is explained by its linear relationship with the spending per pupil?

(f) Interpret the residual plot. What do you see?

(g) What is the correlation between the residuals and spending?

8. A study on the average daily gas consumption for each month and the average number of heating degree-days per day during the month. Here is the regression analysis from Stata.

```
. regress  gasconsumption degreedays

      Source |       SS       df       MS              Number of obs =       9
-------------+------------------------------           F(  1,     7) =  311.97
       Model |  58.9071348     1  58.9071348           Prob > F      =  0.0000
    Residual |  1.32175382     7  .188821975           R-squared     =  0.9781
-------------+------------------------------           Adj R-squared =  0.9749
       Total |  60.2288886     8  7.52861108           Root MSE      =  .43454

------------------------------------------------------------------------------
gasconsump~n |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
  degreedays |   .2022115   .0114485    17.66   0.000     .1751401    .2292829
       _cons |   1.232354   .2860369     4.31   0.004     .5559843    1.908724
------------------------------------------------------------------------------
```

(a) Find the equation of the least-squares line.

(b) Test the null hypothesis that the slope is zero and describe your conclusion.

(c) Give a 99% CI for the slope.

(d) Give a 90% CI for the intercept.

(e) What is the t statistic for testing $H_0 : \beta_0 = 0$?

(f) For the alternative $H_1 : \beta_0 > 0$, how would you report the P-value for this test? Do you reject H_0 at the 5% level?

(g) What is the correlation between daily gas consumption and the average number of heating degree-days?

9. An experiment was conducted involving 12 subjects to study the effect of the percentage of a certain drug in the bloodstream on the length of time (in seconds) it takes to react to a stimulus. Here is the Stata regression output:

```
      Source |       SS       df       MS              Number of obs =      12
-------------+------------------------------           F(  1,    10) =   75.05
       Model |  14.4295851     1  14.4295851           Prob > F      =  0.0000
    Residual |   1.9227069    10   .19227069           R-squared     =  0.8824
-------------+------------------------------           Adj R-squared =  0.8707
       Total |   16.352292    11    1.486572           Root MSE      =  .43849

------------------------------------------------------------------------------
        Time |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      Amount |   .6353147   .0733362     8.66   0.000     .4719114     .798718
       _cons |   .0217366   .3027433     0.07   0.944    -.6528176    .6962907
------------------------------------------------------------------------------
```

(a) What should we do before we ran a regression using Stata?

(b) What is the equation of the least squares regression line?

(c) In the Stata output, in the row labeled Amount, there is a T value of 8.66, and a p-value of 0.000. What null and alternative hypotheses are for this p-value? State your decision for this test using a 1% significance level.

(d) Give the 95% CI for the true slope.

(e) What proportion of variation in reaction times can be explained by the linear relationship between reaction time and amount of drug?

(f) What is the correlation between reaction time and amount of drug?

(g) In the simple linear regression model we assume the y-values are normally distributed with means that vary linearly with x, and with same standard deviation σ. What is the estimate of σ here?

(h) Explain briefly how to check the assumptions for a linear regression model.