Review of topics covered

(Week 06 lecture notes)

Wei Q. Deng

Department of Statistical Sciences

August 13th 2018

Final exam review

- Format is similar to the midterm
- Pay attention to each step you are required show to receive full credits
- Will only go over the question if needed.

Some mathematics

- finding the maximum of a continuous function
- if and only if statements
- Notation for the inverse of quantile
  - $\Phi^{-1}(x) = x$
  - $\Phi^{-1}(-x) = -x$
  - $\Phi(x) = 1 - \Phi(-x) = \mathrm{P}(X < x)$
- know how to take derivative with respect to a continuous function
  - $log(x)$
  - x^k
  - e^x and e^{ax}

Formula provided

- Mean and variance of $Y \sim \chi^2(n-1)$ are $n-1$ and $2(n-1)$, respectively.
- Mean and variance of $Y \sim \text{Binom}(n, p)$ are np and $np(1-p)$, respectively.
- Test statistic for a pair t-test:

$$Z_{\text{d}} = \frac{\bar{X}_1 - \bar{X}_2 + (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

- Pooled variance estimate for two samples with variances σ_1^2 and σ_2^2:

$$S_{\text{p}} = \sqrt{((n_1-1)S_1^2 + (n_2-1)S_2^2)/(n_1+n_2-2)}$$

- Test statistic for a two-sample independent t-test

$$T = \frac{\bar{X}_1 - \bar{X}_2 + (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

Formula provided (cont'd)

- Test statistic for Welch two sample t-test

$$T = \frac{\bar{X}_1 - \bar{X}_2 + (\mu_1 - \mu_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

- ANOVA Test statistic for equality of means in k samples:

$$F = \frac{MST}{MSE} = \frac{\sum_{i=1}^{k} n_i(\bar{X}_i - \bar{X})^2/(k-1)}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij} - \bar{X}_i)^2/(N-k)}$$

- Estimated regression coefficients from least squares method:

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

and

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

Formula provided (cont'd)

- Coefficient of determination:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{a} + \hat{b}X_i - \bar{Y})^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

- Sample correlation

$$r_{X,Y} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

- No pdf/pmf needed.

Topics covered so far

- Lecture 1: Graphical display and summary statistics
- Lecture 2: Sampling distribution and (point estimator of) parameter
- Lecture 3: Point estimates and interval estimates
- Lecture 4: Confidence Intervals and tests for equality of means in one and two samples
- Lecture 6: Hypothesis testing, type I error and power
- Lecture 7 and 8: Introduction to linear regression

From lecture 1:

You should know

- the difference between a sample and a population
- **concepts such as parameter, statistic, estimator**
- the types of data
- what appropriately numerical and graphical summary to use for each type of data
- when to use a boxplot and a histogram
- shapes of data (centre, spread, skewness, shape of tail and outliers)
- **quantile range equivalence to a multiple of standard deviation for a normal random variable**
- how to detect non-normality in data (in terms of shape) through boxplot, histogram and quantile-quantile plots

From lecture 2:

You should know

- definition of a random sample and its properties
- how to check if a sample is roughly random given that it was sampled with replacement and sampled without replacement
- **parameters of normal, t- and binomial/Bernoulli distributions**
- definition of a statistic
- (large sample) sampling distributions of common statistic such as mean, variance, and proportion
- **when the sampling distribution is exact and when to invoke the central limit theorem**
- for a normal random sample,
  - $\bar{X}$ and S^2 are independent,
  - distribution of $\bar{X}$ and $(n-1)S^2/\sigma^2$
  - distribution of $\frac{\bar{X}-\mu}{S/\sqrt{n}}$
- **how to find estimators using MoM and MLE (at least for normal)**
- difference between an estimator and an estimate
- **properties of estimators, unbiasedness, sufficiency and consistency**
- **how to show an estimator is unbiased** ($\bar{X}$ and S^2 are unbiased)

From lecture 3:

You should know

- **how to calculate the MSE by calculating the bias and variance of an estimator**
- **the two computational approaches to estimate the variance of an estimator**
- understand why we want to look at interval estimator on top of point estimators
- **how to obtain an interval estimator for μ**
  - when σ^2 is known
  - when σ^2 is unknown
- **how to obtain an interval estimator for the population proportion p**
- whether the CI becomes wide/narrow as you increase/decrease sample size n, confidence level α
- **how to read the Z-table and t-table for quantile values**

From lecture 4:

You should know

- Structure of a hypothesis test
- how to select a test statistic
- how to calculate the rejection region and p-value given the null and alternative
- definition of a p-value
- statistical tests of μ (normal, t, and binomial)
- Statistical tests to compare two samples
  - test for equality of means (three versions of two-sample t-test)
  - assumptions for different t-tests
  - test for equality of variance (F-test)

From lecture 6:

You should know

- Understand the difference between a one-sided and two-sided test
- Able to structure the appropriate null and alternative hypotheses based on the question (one-sample, two-sample, multiple-samples)
- Understand and able to define type I error and power
- Able to interpret p-value in terms of the strength of statistical evidence
- Know the connection between p-value and CI about the parameter (derived using sample data)

From lecture 6 (cont'd):

- Know how to make a decision with respect to a null hypothesis using
  - p-value and α
  - observed test statistic and rejection region at α-level
  - confidence interval about θ (using sample data) and θ_o
- Know how to construct empirical type I error and power using simulations (pseudo-code)
- Given the derived type II error or power,
  - explain how type II error, power, α, n and $\mu' - \mu$ can influence each other
  - I will not ask you to derive these for any test, but you should understand why AND how they work.
  - able to re-construct the graph on page 34 of lecture 6.
  - here is something that might help you visualize these relationships

From lecture 7 and 8:

- able to identify problems that can be addressed using regression
- principles of training and testing and cross-validation
- **how to construct a regression model with unknown parameters**
- assumptions of linear regression models
- **Construct confidence intervals for the true slope**
- Estimating unknown parameters to obtain an estimated model using R/Python
- Use the estimated model to predict a new observation using R/Python

TRUE OR FALSE

1. A pie chart does not provide a very useful graphical representation for a categorical variable with too many levels. True
2. To increase statistical power of a test, we can usually increase the sample size without changing anything else. True
3. If the observed test statistic falls in the rejection region at significance level α, then we can reject the null hypothesis at significance level 2α. True
4. If we rejected the null hypothesis at significance level α in favour of a two-sided alternative, we can reject the null hypothesis at significance level $\alpha/2$ in favour of a one-sided alternative. True
5. A test statistic is a statistic and thus a function of the sample data. True

TRUE OR FALSE (cont'd)

6. Under random sampling, each member of the collected subset has an equal probability of being chosen. True
7. A two-sample t-test can be seen as a special case of a simple linear regression where the predictor is categorical with two levels indicating which of the two samples each observation belongs to. True
8. A categorical variable can be adequately summarized graphically with a histogram. False
9. The significance level is the same thing as the type I error of a test. False
10. The values at the end of the two whiskers are values in the data that were used to construct the boxplot. False

TRUE OR FALSE (cont'd)

11. We randomly select 20 couples and compare the time the husbands and wives spend watching TV. The difference in time spent watching tv should be tested with a two-sample paired t-test. True
12. Symmetric data can only have one peak. False
13. The sample variance of a normal random sample has a known sampling distribution. True
14. For a random sample with finite variance, the sample mean first converges in distribution to a random normal variable and then converges in probability to the true mean as we increase the sample size. True
15. The MLE and method of moment estimator always agree. False

TRUE OR FALSE (cont'd)

16. Suppose you took 100 random samples of size 50 to construct the 95% confidence interval for the true mean, approximately 95 of those 95% CIs would capture the true mean. True/False
17. Statistical significance implies practical significance (e.g. in a random sample of students 50,600 of 100,000 eats breakfast, the average proportion that eat breakfast is significantly different from 50% at $p < 0.0005$.) False
18. Type I error is evaluated under the sampling distribution specified by the null hypothesis. True
19. If P-value $< \alpha$, the $1 - \alpha$ confidence interval will not contain the parameter value under the null hypothesis. True

Short answer questions

1. What graphical tool/tools is/are the most useful for comparing the distribution of two variables?Scatterplot/QQ-plot
2. For a normally distributed random variable, what is the total length of the boxplot (sum of the two whiskers and the box itself)? Give the length as a multiple of the standard deviation σ.roughly 5.4σ
3. If the sample data is skewed to the left, what would the boxplot look like assuming the data were ordered from left to right in value?
4. State the classical central limit theorem assuming finite variance.
5. What is the parameter of a Bernoulli distribution?p where $p \in (0, 1)$

Short answer questions (cont'd)

6. Show $\text{Var}(X) = \text{E}[X^2] - \mu^2$ (Hint: By definition $\text{Var}(X) = \text{E}[(X-\mu)^2]$).
7. What is the sample size needed to invoke central limit theorem?
8. Suppose $X_1, \ldots, X_n \sim \mathcal{N}(0, \sigma^2)$ are a random normal sample, what is the sampling distribution of $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ if σ^2 is unknown? (Define any quantity you need.) $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ where S is the squared root of sample variance or standard error
9. Name two common methods to find an estimator for a sample from the parametric family. Briefly explain how they work. MLE and MoM
10. Write out the likelihood of $x_1, \ldots, x_n$ assuming they were sampled from a normal distribution with parameters mean μ and variance σ^2.

Short answer questions (cont'd)

11. What is the MSE of an estimator $\hat{\theta}$ with respect to the parameter θ? $\mathrm{E}[(\hat{\theta} - \theta)^2] = \mathrm{Var}(\hat{\theta}) + (\mathrm{E}[\hat{\theta} - \theta])^2$.
12. What is a type I error? Give an example of a type I error.
13. What is a type II error? Give an example of a type II
14. What is the statistical technique behind bootstrap and jackknife methods to estimate standard errors? random sampling
15. Name one measure of association between two variables.Pearson's correlation coefficient or Coefficient of determination

Short answer questions (cont'd)

16. For a random sample $X_1, \ldots, X_n$ with mean μ, given an approximated 95% confidence interval for μ when σ^2 is known.
17. For a random sample $X_1, \ldots, X_n$ with mean μ, given an approximated 95% confidence interval for μ when σ^2 is NOT known.
18. For a random sample $X_1, \ldots, X_n \sim \text{Ber}(p)$, given the approximated 95% confidence interval for p.
19. For a given $1 - \alpha$ confidence interval (CI) for μ, the population mean, name two ways to decrease the width of that CI.
20. Let $x_1, \ldots, x_n$ be an observed random sample from a population. Write out the pseudo code to obtain an estimate of standard error for the 75% percentile $\tilde{x}_{0.75}$? non-parametric boostrap - make sure you know each step.

Short answer questions (cont'd)

21. Name at least two assumptions for two-sample independent t-test.
22. Name at least two assumptions for two-sample paired t-test.
23. What is the purpose of a statistical test?
24. Suppose $X_1, \ldots, X_n \sim \mathcal{N}(\mu, 5)$, where n $= 100$. We want to test the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$). Give the procedure to evaluate the empirical type I error if we were to repeat the testing 1000 times.
25. Suppose $X_1, \ldots, X_n \sim \mathcal{N}(1, 5)$, where n $= 100$. We want to test the null hypothesis that the true mean $\mu = 0$ with a two-sided test (i.e. the alternative is $\mu \neq 0$). Give the procedure to evaluate the empirical power to detect $\mu = 1$ if we were to repeat the testing 1000 times.

Derivations

1. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Define $T(X_1, \ldots, X_n) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $T(X_1, \ldots, X_n) = S^2 = \frac{1}{n-1}\left[\sum_{i=1}^{n} X_i^2 - n\bar{X}^2\right]$. Show

(a) $\mathrm{E}(\bar{X}) = \mu$

(b) $\mathrm{Var}(\bar{X}) = \frac{\sigma^2}{n}$

(c) $\mathrm{E}(S^2) = \sigma^2$

Derivations

2. Let $X_1, \ldots, X_n$ be a random sample from a population with mean μ and variance $\sigma^2 < \infty$. Define $T'(X_1, \ldots, X_n) = \tilde{\theta} = \frac{1}{n} \sum_{i=1}^{n} \frac{n}{n-1} X_i$. Show

(a) $\mathsf{E}(\tilde{\theta}) = \frac{n}{n-1}\mu$

(b) $\mathsf{Var}(\tilde{\theta}) = \frac{n\sigma^2}{(n-1)^2}$

Derivations (cont'd)

3. Construct a 95% CI for μ assuming σ is known.
4. Construct a 95% CI for μ assuming σ is NOT known.

Derivations (cont'd)

5. Show that $\hat{a}$ and $\hat{b}$ are the least square solution to minimizing $\sum_{i=1}^{n}(y_i - bx_i - a)^2$ where the outcome variable is $y = (y_1, \ldots, y_n)$ and the predictor is $x = (x_1, \ldots, x_n)$.

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

and

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

Derivations (cont'd)

6. Calculate the coefficient of determination of a simple linear model $y \sim a + bx$ as the squared correlation between the outcome $y = (y_1, \ldots, y_n)$ and the predictor $x = (x_1, \ldots, x_n)$. The coefficient of determination is:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{a} + \hat{b}x_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

and the Pearson's correlation coefficient is:

$$r^2 = \frac{(\sum(x_i - \bar{X})(y_i - \bar{x}))^2}{[\sum(x_i - \bar{x})^2][\sum(y_i - \bar{y})^2]}$$

(Hint: write $\hat{a} = \bar{y} - \hat{b}\bar{x}$)

Solving problems

1. In the sample of 400 students, the sample mean of student heights is $x_{\text{height}} = 1.70$ and the sample standard deviation is $s_{\text{height}} = 0.088$ meters. In this case, we can confirm that the observations are independent by checking that the data are a random sample drawn from the population.

(a) Give the sampling distribution of the sample mean of student heights. $\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$ since the standard error is given

(b) What is the 95% confidence interval for the true average height of the population?

(c) What is the 95% confidence interval for the true average height of the population? See #16 of T/F questions

(d) Would you be surprised if someone told you the average height of the population was actually 1.69 meters? Roughly what is the probability of expecting the true mean to be 1.69 or lower under the sampling distribution?

Solving problems (cont'd)

(e) What are the null and alternative hypotheses if you would like to test whether the average height of the population is actually 1.69 meters? Define μ as the population mean first.

(f) Evaluate the hypotheses by calculating the observed test statistic and p-value.

(g) Interpret the results and make a concluding statement.

(h) Suppose everything remains the same, but the sample size was actually incorrectly recorded as 100 instead of 400. How would you conclusions change and why? Why has the conclusion changed? Relate to the concept of statistical power - the ability to detect a difference when there is truly a difference.

Solving problems (cont'd)

2. Colleges frequently provide estimates of student expenses such as housing. A consultant hired by a community college claimed that the average student housing expense was \$650 per month.

(a) What are the null and alternative hypotheses to test whether this claim is accurate?

(b) Identify the test statistic and give the sampling distribution of the test statistic. Since σ is not given, assume it is unknown and define S, additionally assume sample size is greater than 30.

(c) The sample mean for student housing is \$616.91 and the sample standard deviation is \$128.65. Construct a 95% confidence interval for the population mean.$(\bar{X} - t_{0.975,n-1}S/\sqrt{n}, \bar{X} + t_{0.975,n-1}S/\sqrt{n})$

(d) Evaluate the hypotheses using the 95% confidence interval (i.e. the significance level is 5%). **Change the question to: what is the sample size needed such that the 95% CI covers the null hypothesis parameter?**$\bar{X} + t_{0.975,n-1}S/\sqrt{n} > 650$; $n \sim 58$; check the CI and see if it contains 650).

Solving problems (cont'd)

(e) What would the null and alternative hypotheses be if the university would like to assess whether an initiative for low-cost housing for students is plausible?

(f) Construct a lower one-sided as well as an upper one-sided 95% confidence interval for the population mean.$(\bar{X} + t_{0.95,n-1}S/\sqrt{n}, \infty)$, $(-\infty, \bar{X} - t_{0.95,n-1}S/\sqrt{n})$

(g) Use one of the one-sided intervals you constructed above to make a conclude with respect to the hypothese in e) **using the smallest sample size from d)**.$(\infty, 588.4)$ and $(645.4, \infty)$

Solving problems (cont'd)

3. A manufacturer of textbooks is interested in estimating the strength of the bindings produced by a particular machine. Strength is measured by the force required to pull the pages from the binding. A random sample of recorded pound-force is given by $(X_1, \ldots, X_n)$.

(a) Give the 95% confidence interval for the average force required to break (μ) assume $\sigma = 0.8$.

(b) If the force is measured in pounds, how many books should be tested to estimate the average force required to break the binding within .1lb with 95%?The question asks for the sample size required such that $\mu \pm 0.1$ should be covered in the 95% confidence interval.

(c) Suppose now $\sigma = 0.5$, what sample size is required to achieve a 95% CI with width 0.1?

(d) The observed sample mean force is 2.2 lb. What is the minimum sample size required such that the 95% CI covers $\mu' = 3$ assuming $\sigma = 0.5$?

Solving problems (cont'd)

4. Your instructor is sloppy when typing in Latex and often makes typos on the lecture slides. One diligent student records whether each page contains a typo or not during class and 60 out of the 500 pages had typos. Suppose the status of a typo in each page of the lecture slides follows a Bernoulli distribution with probability p, the students in the class are wondering whether the proportion of pages with typos is less than 30%.

(a) Identify the parameter of interest in mathematical symbols and state the null and alternative hypothesis.

(b) What is the test statistic and its sampling distribution under the null hypothesis?

(c) Using the data, calculate a 95% confidence interval (CI) assuming the variance of the estimator can be adequately approximated by the sample variance. Does this interval contain the parameter value under the null hypothesis?

Solving problems (cont'd)

(d) Calculate the test statistic and compare it with the quantile of the sampling distribution at $\alpha = 0.05$.

(e) Suppose the sample proportion is unchanged, but the instructor made more slides so $n = 1000$. Repeat the calculation of test statistic and compare it with the quantile of the sampling distribution at $\alpha = 0.05$.

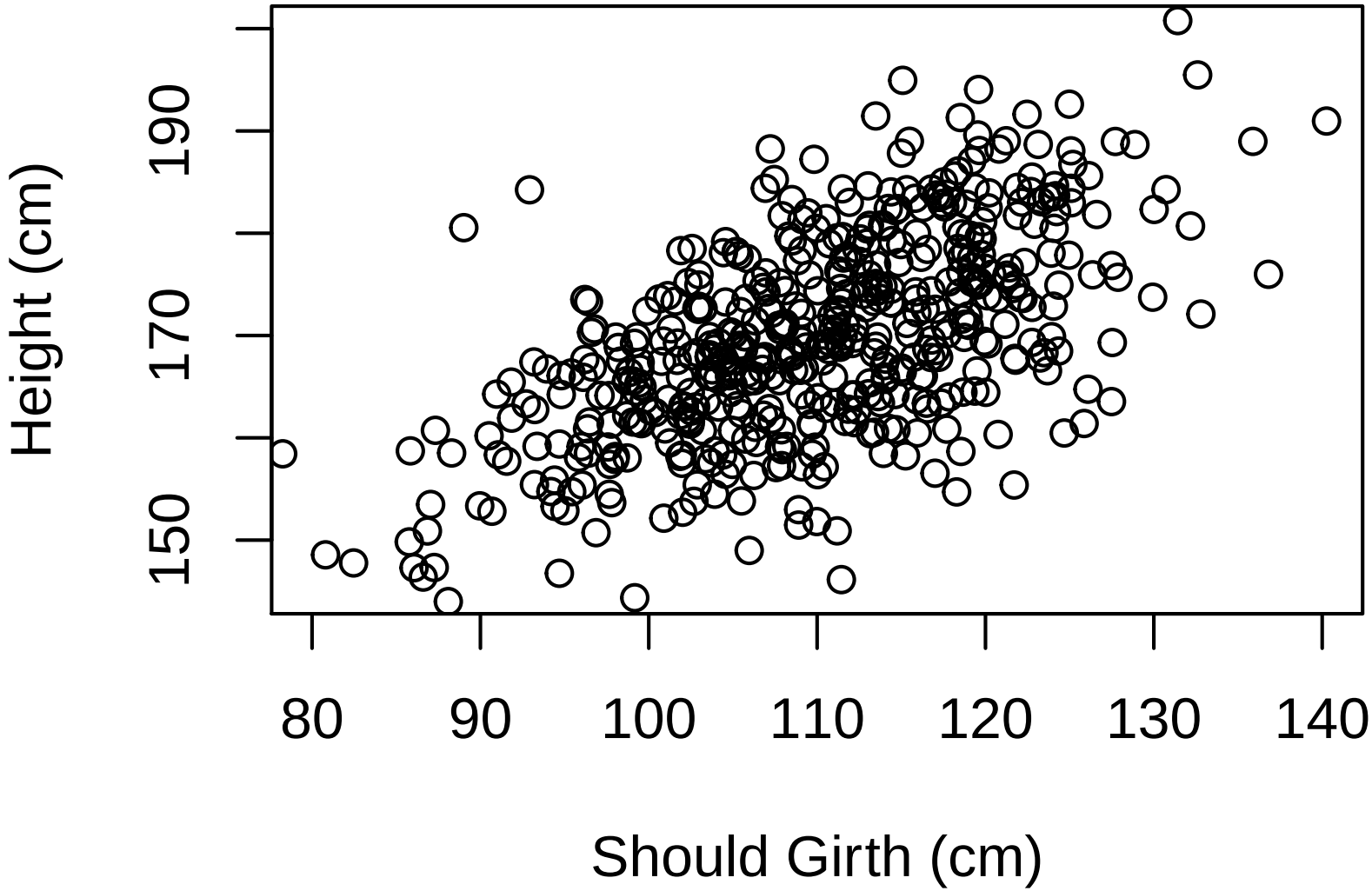
(f) Do your conclusions in (c) and (d) agree? Why or why not?

(g) Does your conclusion in (c) or (d) change if the alternative hypothesis is two-sided?

(h) What changes can you make to increase the statistical power of this test?

Solving problems (cont'd)

5. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (over deltoid muscles), both measured in centimeters.



Solving problems (cont'd)

(a) Describe the relationship between shoulder girth and height.

(b) Show your answer to the question above mathematically.

(c) If the R^2 for the least squares regression line for the data is 36%, what is the correlation between height and shoulder girth? What is the intepretation in this context?

(d) The mean shoulder girth is 110 cm with a standard deviation of 10 cm. The mean height is 170 cm with a standard deviation of 10 cm. Write the equation of the regression line for predicting height.

(e) Interpret the slope and the intercept in this context.

(f) A randomly selected student from your class has a shoulder girth of 100 cm. Predict the height of this student using the model.

(g) The student's actual height is 160 cm tall. Calculate the residual, and explain what this residual means.

(h) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child?

Solving problems (cont'd)

(j) Construct a 95% CI for the true slope.

(k) Is the true slope significantly different from 0? Write the null and alternative hypotheses.

(l) Compute the p-value for the hypothesis test. Does the conclusion agree with the one based on the 95% CI?

(m) Calculate the slope and intercept given that shoulder girth and height are now measured in inches.

(n) What do you conclude from the results above?