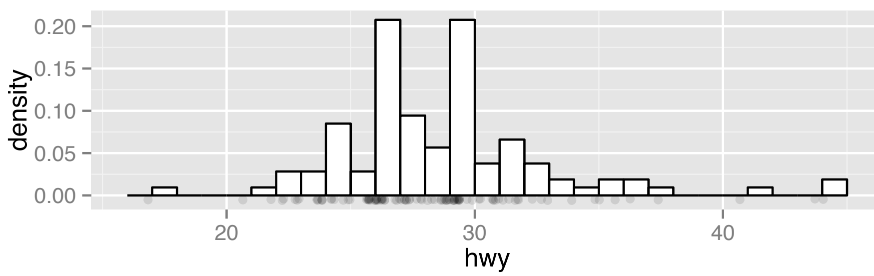
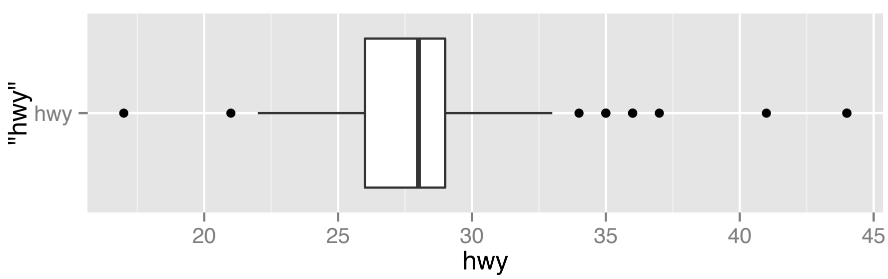
Advanced Data Analysis 1, Stat 427/527 Midterm Chapters 1–5 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

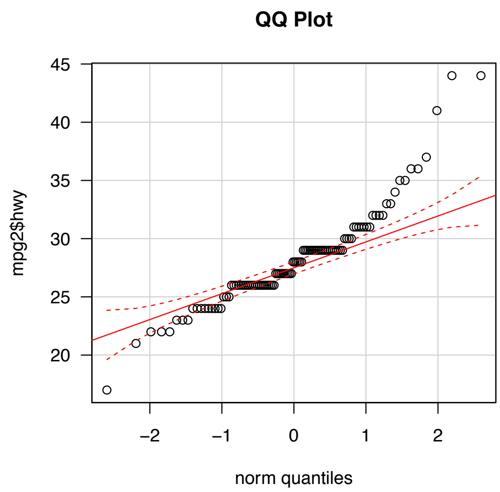
Part I. (250 points) The exam has three problems. Statistical results are presented first, then you are asked to interpret the results. Try to be thorough yet succinct with your answers. I recommend reading through all the questions before you begin writing and following the directions carefully.

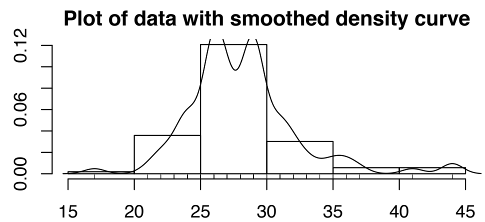
(140pts) **1. Four-wheel drive highway miles per gallon:** The `mpg` dataset is included in the `ggplot2` package. "This dataset contains a subset of the fuel economy data that the EPA makes available on http://fueleconomy.gov. It contains only models which had a new release every year between 1999 and 2008 — this was used as a proxy for the popularity of the car." We study the highway miles per gallon for front-wheel drive models. You may find some of the output below helpful for answering the following questions.

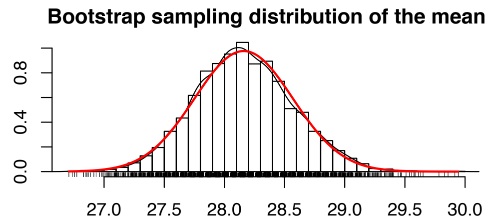





```
##
##  One Sample t-test
##
## data:  mpg2$hwy
## t = 0.3925, df = 105, p-value = 0.6955
## alternative hypothesis: true mean is not equal to 28
## 95 percent confidence interval:
##  27.35 28.97
## sample estimates:
## mean of x
##     28.16
##
##  Anderson-Darling normality test
##
## data:  mpg2$hwy
## A = 2.715, p-value = 6.929e-07
```


QQ Plot
mpg2$hwy
norm quantiles


Plot of data with smoothed density curve


Bootstrap sampling distribution of the mean

(a) (10 pts) *Using only the boxplot from the hwy data, answer the questions (a)–(d):*
What is the shape of the distribution? Explain in detail what in the boxplot tells you this is the shape.

Solution:
A few more extreme high values than low values.
Generally symmetric, or very weak evidence for right-skew (-1 pt).

(b) (10 pts) What, approximately, are the values of the following?

Minimum:__________ Q1:__________ Median:__________

Q3:__________ Maximum:__________ IQR:__________

Solution:
From the boxplot:

```
fivenum(mpg2$hwy)
## [1] 17 26 28 29 44
IQR(mpg2$hwy)
## [1] 3
```

(c) (10 pts) Will the mean be larger or smaller than the median? Explain.

Solution:
$\bar{Y}$ roughly equal to M because of symmetry, possibly slightly larger because of the few high outliers.

```
mean(mpg2$hwy)
## [1] 28.16
median(mpg2$hwy)
## [1] 28
```

(d) (10 pts) If I needed to construct a 95% CI for μ, would I need to be very concerned about using the one-sample t-procedure to do so? Explain.

Solution:
The sample size is large, data are symmetric, and there are only a couple outliers which are not very extreme, so there is not much to worry about.
OR
Because of the few large outliers, we should worry a little. (Good practice would follow up the t-procedure with a nonparametric procedure.)

(e) (10 pts) *Now examine the histogram.*
What features of the distribution do you see in the histogram that you do not see in the boxplot?

Solution:
Unimodal or Bi-modal; there are certainly two clustesr of points at 26 and 29.

(f) (10 pts) If I were to compute a CI for μ using the data in the plots, what would be the population for which I was estimating μ, and how would I interpret μ?

Solution:
Population is four wheel cars with models released between 1999 and 2008. (Our sample are a subset of cars released every of those years, the "popular" cars.)
μ is the population mean highway miles for four wheel cars with models released between 1999 and 2008.

(g) (10 pts) Examine the output from the 1-sample t-procedure. What null and alternative hypotheses are being tested? State them carefully.

Solution:
$H_0 : \mu = 28.0$ versus $H_A : \mu \neq 28.0$ (in mpg).

(h) (10 pts) What is the p-value for the test in part (g)? Explain carefully what this value means.

Solution:
$p = 0.695$.
This is the probability of observing a sample mean (actually the t-statistic) at least as extreme (that is, further away from $\mu_0 = 28.0$) as the one observed conditional on H_0 being true.
Saying that the p-value is a measure of plausibility of H_0 is less clear.

(i) (10 pts) With respect to the hypotheses in part (g), what would you conclude using a level $\alpha = 0.05$ test? Why do you conclude this?

Solution:
Since $p = 0.695 \geq 0.05$, we fail to reject H_0 concluding that either that the data are consistent with H_0, or that we have insufficient evidence to conclude $\mu \neq 28.0$.

(j) (10 pts) Of the two types of error, Type-I or Type-II, which could you be committing in part (i)? Explain. Do you know if you are committing this type of error?

Solution:
A Type-I error is rejecting H_0 when it is true.
A Type-II error is failing to reject H_0 when it is false.
Since we failed to reject H_0, we could only make a Type-II error here, but it is unknown to us whether such an error was made.

(k) (10 pts) What is the confidence interval reported from the 1-sample t-procedure? Give a complete interpretation of this interval.

Solution:
The 95% CI is $27.4 \leq \mu \leq 29$. We are 95% confident the population mean miles per gallon for front-wheel drive cars is between 27.4 and 29 (mpg).

(l) (10 pts) If you were to test H_0: $\mu = 30$ vs. H_A: $\mu \neq 30$, what could you conclude? Explain, and be complete.

Solution:
Since $\mu_0 = 30$ is not contained in the 95% CI for μ, we would reject $H_0 : \mu = 30$ at the $\alpha = 0.05$ significance level in favor of H_A.

(m) (10 pts) Look at the probability plot and normality test of hwy, as well as the bootstrap sampling distribution of the sample mean. Interpret each one, including what formal test is being conducted (related to a hypothesis test on the previous page).

Solution:
The shape (S-shaped), of the normal probability plot is consistent with a distribution having long tails (or U-shaped, being slightly right-skewed).
The formal (Anderson-Darling) test of normality has a p-value < 0.005, so the data do not appear to be from a population having a normal distribution.
The bootstrap sampling distribution of the mean indicates a normal distribution.

(n) (10 pts) Is there indication of a problem with the t-test and CI? Explain.

Solution:
Therefore, departures from normality from the original sample is unlikely to cause an important problem with using the t-CI and t-test.

(20^{pts}) **2. Pulse rates:** Students in an introductory statistics class (MS212 taught by Professor John Eccleston and Dr Richard Wilson at The University of Queensland) participated in a simple experiment. The students took their own pulse rate. They were then asked to flip a coin. If the coin came up heads, they were to run in place for one minute. Otherwise they sat for one minute. Then everyone took their pulse again.

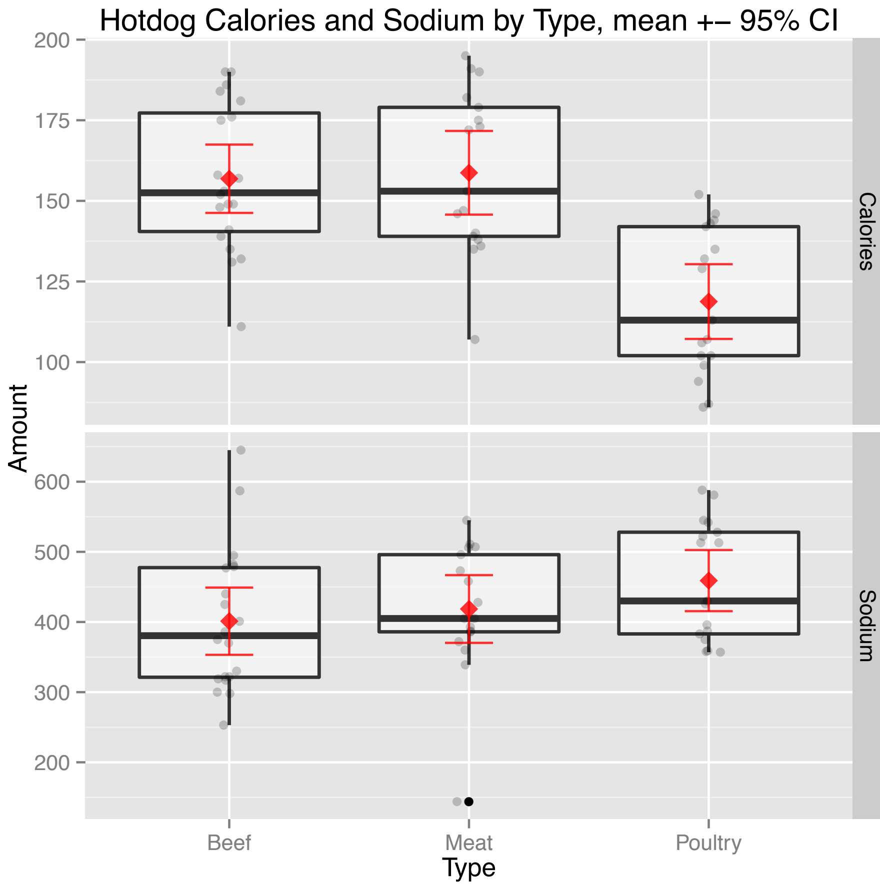
(a) (10 pts) Explain whether the two measurements for each student who ran in place must be treated as paired or as independent, and why.

Solution:
The two measurements for each person who ran are paired since they were both taken on the same individual.

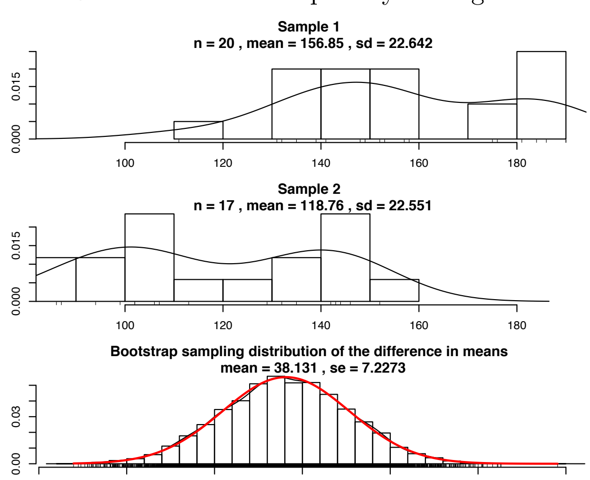
(b) (10 pts) Explain whether the data for the two groups of students (those who ran in place vs those who did not run) are paired or independent, and why.

Solution:
The two groups are independent since the measurements from a person in one group is unrelated to a measurement from a person in the other group.

(90pts) **3. Hot dogs:** People who are concerned about their health may prefer hot dogs that are low in salt and calories. Our data are the results of a laboratory analysis of sodium and calories contained in three types of hot dog: beef, poultry, and meat (mostly pork and beef, but up to 15% poultry meat), in each of 54 major hot dog brands (Consumer Reports, June 1986, pp. 366-367). Varibles included are `Type` of hotdog (beef, meat, or poultry); `Calories` per hot dog; and `Sodium` in milligrams per hot dog. Below are plots, a two-sample t-test, and ANOVA with multiple comparisons.
Plot of the data for this question:



Two-sample bootstrap plot for Calories of beef and poultry hotdogs:



Two-sample t-test for Calories of beef and poultry hotdogs:

```
##
##  Welch Two Sample t-test
##
## data:  calories.beef and calories.poultry
## t = 5.11, df = 34.09, p-value = 1.229e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  22.94 53.23
## sample estimates:
## mean of x mean of y
##     156.8     118.8
```

Means for Sodium by type of hotdog:

```
##      Type     m
## 1    Beef 401.1
## 2    Meat 418.5
## 3 Poultry 459.0
```

ANOVA for Sodium by type of hotdog:

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## Type         2  31739   15869    1.78   0.18
## Residuals   51 455249    8926
```

Multiple comparisons:

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  hotdog$Sodium and hotdog$Type
##
##         Beef  Meat
## Meat    0.580 -
## Poultry 0.069 0.217
##
## P value adjustment method: none
```

```
##   Tukey multiple comparisons of means
##     90% family-wise confidence level
##
## Fit: aov(formula = Sodium ~ Type, data = hotdog)
##
## $Type
##               diff     lwr    upr  p adj
## Meat-Beef    17.38 -48.055  82.81 0.8430
## Poultry-Beef 57.85  -7.584 123.28 0.1620
## Poultry-Meat 40.47 -27.564 108.51 0.4304
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  hotdog$Sodium and hotdog$Type
##
##         Beef Meat
## Meat    1.00 -
## Poultry 0.21 0.65
##
## P value adjustment method: bonferroni
```

(a) (10 pts) *Examine the two-sample t output on the Calories example.*
State the null and alternative hypotheses being tested.

Solution:
$H_0 : \mu_b = \mu_p$ or $\mu_b - \mu_p = 0$ versus
$H_A : \mu_b \neq \mu_p$ or $\mu_b - \mu_p \neq 0$,
where μ_b and μ_p are the population mean calories amounts for beef and poultry hotdogs, respectively.

(b) (10 pts) What do you conclude about the test in part (a) using an $\alpha = 0.05$ significance level? Explain.

Solution:
p-value for test is 1.23×10^{-5}. For $\alpha = 0.05$ significance level, because $p = 1.23 \times 10^{-5} < 0.05 = \alpha$, we reject H_0 in favor of H_A concluding the amount of calories is significantly different between beef and poultry hotdogs.

(c) (10 pts) What assumption is being made about variances in the two-sample t-test? Be specific about this output, i.e., is this the pooled or Satterthwaite form?

Solution:
This is Satterthwaite's t-test, which does not assume equal population variances. We know this because the $df \neq n_1 + n_2 - 2$ (and pooled df does equal this).

(d) (10 pts) What sampling distribution is relevant here? What assumption is being made about that sampling distribution? Does this assumption appear to hold?

Solution:
The we assume the sampling distribution of the difference in means, $d = \bar{y}_1 - \bar{y}_2$, is normal. This assumption appears satisfied (or not contradicted).

(e) (10 pts) What is the CI from the two sample t-test output? Interpret this carefully. How is it consistent with the test of hypothesis?

Solution:
The 95% CI for $\mu_b - \mu_p$ has limits 22.9 to 53.2. We are 95% confident that the population mean calories amount for beef (μ_b) is 22.9 and 53.2 greater than the mean calories for poultry (μ_p) hotdogs. The CI does not contain 0 as a plausible value for $\mu_b - \mu_p$ which is consistent with rejecting H_0 : $\mu_b - \mu_p = 0$ at the 0.05 significance level.

(f) (10 pts) *Now examine the one-way ANOVA output on the Sodium example.*
What is the hypothesis being tested by the F-test?

Solution:
$H_0 : \mu_b = \mu_m = \mu_p$, where μ_j $(j = b, m, p)$ is the population mean for beef, meat, and poultry hotdogs, versus
H_A : not H_0, at least one pair of means differ.

(g) (10 pts) Define the populations for which you are testing hypotheses.

Solution:
Population j is hotdogs of type $(j = b, m, p)$ from major brands,

(h) (10 pts) What can you conclude from the F-test alone using an $\alpha = 0.10$ significance level? Explain why, and be complete.

Solution:
The p-value for F-test is 0.1793, so we fail to reject H_0 at the $\alpha = 0.10$ significance level, concluding we have insufficient evidence that any population means differ from each other.

(i) (10 pts) Regardless of the decision made in part (g), report the Tukey groupings of means, and explain how you obtained these.

Solution:
μ_b and μ_m not sig different since 0 in CI (or p-value< 0.10).
μ_b and μ_p not sig different since 0 in CI (or p-value< 0.10).
μ_m and μ_p not sig different since 0 in CI (or p-value< 0.10).
This leads to these groupings, increasing mean sodium:

```
 b   m   p
-----------
```