

June 14, 2010

1. Here's some code... know what's what:

```
#FIXED X'S
X=seq(10,20,.25) Xt2=X*X Xt.5=sqrt(X) n=length(X)

#SIMULATED REGRESSION MODEL
alpha=1 beta=.5 sigma=.5 Y=alpha+beta*X+rnorm(n, 0, sigma) plot(X,Y)

#DATA STORAGE
my.data=data.frame(Y,X,Xt2,Xt.5) names(my.data)

#REGRESSION MODEL FITTING
lm.fit=lm(Y~X,data=my.data) lm.fit summary(lm.fit)

#REGRESSION FITTING OUTPUT
names(lm.fit) lm.fit$coefficients lm.fit$residuals lm.fit$fitted.values

#RESULT
plot(X,Y) abline(lm.fit$coefficients[1],lm.fit$coefficients[2])

#DIAGNOSTICS
plot(lm.fit$fitted.values,Y) plot(lm.fit$fitted.values,X)
plot(lm.fit$fitted.values,lm.fit$residuals) plot(X,lm.fit$residuals)

#PREDICTION, for E[Y|X] and Y|X, respectively
predict(lm.fit,newdata=data.frame(X=c(15,20)), interval = "confidence")
predict(lm.fit,newdata=data.frame(X=c(15,20)), interval = "prediction")
```

2. Demonstrate the uncertainty in fitting regression models (simulation).
3. Fit a simple linear model to your data (simple means a single covariate X , like above).
4. Estimate some coefficients (there's uncertainty, right?).
5. Examine the regression assumptions in your context (hmm... how will you do this?).
6. Do some prediction (there's uncertainty, right?),
7. and then give intervals for where new Y 's might be for certain levels of X .
8. Hint/Question for previous two questions: What's the difference between the two predictions above? (R seems to call one 'confidence' and one 'prediction')?
9. Last: Do X 's that are farther apart allow for better model estimation? I.e., reduced uncertainty in model fit? Show this theoretically as well as through simulation.
10. Last last: 12-5, 12-8, 12-9, 12-18

June 11, 2010

- 11-2
- 11-3 (Be careful, the book defines x and y a funny way)
- 11-8
- Suppose the regression (normal) model,

$$Y_i = \alpha + \beta X_i + \epsilon_i, \quad \epsilon_i \stackrel{iid}{\sim} N(0, \sigma),$$

and let $\hat{Y}_i = a + bX_i$ be a least squares fit for the observed data, so that $Y_i = a + bX_i + e_i$.

- Prove $b \sim N\left(\beta, \sqrt{\sigma^2 / \sum (X_i - \bar{X})^2}\right)$
- Prove $a \sim N\left(\alpha, \sqrt{\sigma^2/n + (\bar{x}^2 \sigma^2) / \sum (X_i - \bar{X})^2}\right)$

For a new $X_{new} = x_0$, how might we predict Y_{new} ? Assuming you got this correct...

- Prove $\Pr(\hat{Y}_{new} | X_{new} = x_0) = N\left(\alpha + \beta X_i, \sqrt{\sigma^2/n + (x_0 - \bar{X})^2 \sigma^2 / \sum (X_i - \bar{X})^2}\right)$

If you observe that $\Pr(e_i) \approx N(0, s)$, what is $\Pr(Y_{new} | X_{new} = x_0)$?

What is a good estimator for σ ? (You may want to check the book to make sure about this).

- 1-18 (remember?) What does 1-18 have to do with regression?
- 3-46... extra, for fun.

June 7, 2010

- 7-2
- 7-3
- 7-4
- 7-5
- 7-6
- 7-8
- 7-9
- 7-11
- That's it.
- Make sure you're ready for the test

June 2, 2010

1. Choose a null hypothesis $E[X] = \mu_{H_0}$ for problem 1 of the last homework. Test your null hypothesis at the $\alpha = .10$ significance level using a classical testing procedure (i.e., without the use of confidence intervals). What do you conclude?
2. Return to problem 3 on the previous homework. Test $H_0 : E[X] = E[V]$ at the $\alpha = .01$ significance level using a classical testing procedure (i.e., without the use of confidence intervals). What do you conclude? You have made two major assumptions in carrying out this test – one involving the CLT and one involving the population variances – what are your assumptions?
3. Prove that if some μ_{H_0} is not in a 95% confidence interval, then it will be rejected by a classical two-sided test at the $\alpha = .05$ level. Do you conclude that $(1 - \alpha)\%$ confidence intervals are equivalent to two-sided α level tests?
4. Busy work: For each of the book problems on the last homework assignment, what was the set of plausible (e.g., those that would not be rejected at the $\alpha = .05$ significance level) hypothesis found?
5. Book problems... second to last major problem set for the last test...
 - 9-4
 - 9-5
 - 9-7
 - 9-8
 - 9-9
 - 9-12
 - 9-13
 - 9-14
 - 9-15
 - 9-16
 - 9-18
 - 9-23
 - 9-24
 - 9-25
 - 9-26
6. A few old problems that are good to do...
 - 7-3
 - 7-7
 - 7-10
 - 7-15

June 1, 2010

1. Give a 90% confidence interval for the population mean $E[X]$ of a variable X in your data. Interpret the interval – both in its statistical properties, and in its meaning for your data.
2. What are the degrees of freedom (df) associated with a confidence interval for the difference of two population means (i.e., $E[X] - E[Y]$) when we do not assume $\text{Var}[X] = \text{Var}[Y]$? WIKI!
3. Divide your variable X from problem 1 into two groups on the basis of some other relevant feature in your data, i.e., make one group the ‘treated’ group and the other the ‘control’ group. What is the ‘treatment’? Is there a potential for confounding, i.e., are there differences between the treated and control groups other than the treatment? Give a 99% confidence interval for the difference in population means of the treated group and the control group. You have made two major assumptions in creating this interval – one involving the CLT and one involving the population variances – what are your assumptions?
4. Suppose $X_i \sim \text{binomial}(N, p)$. What does \bar{x} estimate? In a usual setting, do you think you will need to estimate N , or will it be known? How will you estimate p ? Provide two ways to estimate $\text{Var}[X]$ from a sample X_1, \dots, X_n . Hint: What is $\text{Var}[X]$ in this setting?
5. Book problems... like what you might could have to do for the test...
 - 8-2
 - 8-3
 - 8-4
 - 8-6
 - 8-7
 - 8-8
 - 8-10
 - 8-11
 - 8-13
 - 8-15

For the remaining binomial-based problems, the population variance is replaced with a sample-derived estimate, but (for simplicity) we will assume that we *do not* need to change from a normal distribution to a t distribution as a result of the approximation.

- 8-17
- 8-19
- 8-21

May 28, 2010 (4 point assignment!)

- Using the **sample** function, make your own crazy discrete distribution to sample from. You could actually use some of your data... pretending it was a whole population...
 - Show me your distribution.
 - Show me random samples from your distribution for sample sizes of $n = 5$, $n = 10$, $n = 30$, $n = 50$, $n = 100$, $n = 200$, $n = 500$, and $n = 1000$. Do the observed empirical distributions better approximate your distribution as n grows?
 - Now, for each n above, for a large number of repetitions r , repeatedly resample random samples of size n from your distribution. Calculate the sample average for each of the r repeated samples – this gives you r sample averages – show me a histogram of these values and report the corresponding sample mean and sample standard deviation. I.e., start with $n = 5$, take r random samples from your distribution, and make a histogram and calculate the average and standard deviation of these values; then set $n = 10$ and do the same, etc.
 - For a random variable X sampled from your distribution, what $E[X]$ and $\text{Var}[X]$?
 - For each n above, what is the approximate distribution of \bar{x} according to the central limit theorem (CLT) for random samples from your distribution?
 - Does this match your simulation results? Be specific about when it *does not*.
 - Produce some nice reusable code that creates a ‘movie’ demonstration of the CLT. I will visit with each of you individually in the following lab to watch your demonstration. The code should be general enough so we can run it for different sample sizes n , and can change the distribution we are sampling from easily. *This should be easy to do based on the early work from this problem.* What I have in mind is, the movie shows (one at a time – in sequence) r repeated samples from your distribution in one figure, and a histogram of all the averages so far in another figure.

```
#Useful functions for this problem:
sample #define your own distribution (discrete)
par(mfrow=c(1,2)) #put two plots in the same window
readline() #wait for keyboard input
#and don't forget about the for loop... and the other usual stuff...
```

- In a previous homework problem you considered $E[X]$ and $\text{Var}[X]$ for a binomial random variable. For a random sample of size X_1, \dots, X_n , what is the distribution of \bar{x} according to the CLT? You can copy the answer from the book, but if it doesn't make sense to you, you're going to be in trouble later...
- How does the previous problem apply to predicting presidential elections?
- To correctly answer this problem, write in your own hand writing exactly what I have written here: “The sampling distribution of \bar{x} for a random random sample of size n from some distribution P will be approximately normal according to the central limit theorem (CLT). The approximation gets better as n gets bigger, and for reasonably shaped, symmetric distributions, $n = 30$ will likely be sufficient for good approximation.”

May 25, 2010

1. Book problems. Copying answers from the back of the book will not help you on the test. Checking to make sure you've done problem correctly probably will.
 - 1-9
 - 1-11
 - 1-14
 - 1-18
2. Describe the design of your data in a paragraph or two:
 - Is the data observational, or from a controlled experiment?
 - Do you have a sample, or a population?
 - Envision your data as a sample from a population: What population is it from?
 - Are there any variables NOT in your data, that might associate with your variables?
 - Do you have variables in your data that can be interpreted as 'treatments'?
 - What variables don't you have that could be confounders relative to your 'treatments'?
3. More book work.
 - 7-2
 - 7-6
 - 7-8
 - 7-9
 - 7-14 a-c
 - 7-16
 - 7-17
4. Let X_1, \dots, X_n be a sample from a population with some parameter θ . Let F be an estimator of θ . Let f be a realized estimate of F for θ based on observed x_1, \dots, x_n . F has a distribution since it is calculated from the sample X_1, \dots, X_n , e.g., a different sample produces a different estimate f , i.e., a different sample from the distribution of F . By definition, $\text{MSE} = \sum_f (f - \theta)^2 \Pr(f)$. Prove that this is equal to $\text{Var}[F] + (\text{bias}[F])^2$.

May 24, 2010

1. Prove mathematically that $E[aX + bY + c] = aE[X] + bE[Y] + c$ for $(X, Y) \sim \Pr(X, Y)$.
2. For X and Y above, also prove that $V[aX + bY + c] = a^2\text{Var}[X] + b^2\text{Var}[Y] + 2ab\text{Cov}[X, Y]$.
3. Provide sample analogues to the population parameters of covariance and correlation.
4. Show, using your data, that the **cov** and **cor** functions calculate these, respectively.
5. Let $Y_i = \beta X_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. Find values for β and σ^2 that produce sample correlations of approximately -.8, 0, 0.5, 0.9, and 1 between Y and X , and plot the samples.
6. Plot two interesting variables in your data against each other and report their correlation.
7. Covariance and correlation measure LINEAR association only. Illustrate this with an example. Hint: Invent or find a bivariate data set that has covariance equal to 0, but that is in fact highly structured, with an obvious non-linear relationship between its two variables.
8. Book problems. Copying answers from the back of the book will not help you on the test. Checking to make sure you've done problem correctly probably will.
 - 5-5
 - 5-7
 - 5-9
 - 5-13
 - 5-15
 - 5-17

May 21, 2010

1. Show using some of your data that if $y = ax + c$, then $\bar{y} = a\bar{x} + c$.
2. Show using some of your data that $\overline{x^2} = \sum_{i=1}^n x_i^2/n \neq \bar{x}^2$.
3. Show using some of your data that if $y = ax + c$, then $s_y = as_x$.
4. Suppose that the distribution of ounces of water that a person drinks each day is $N(40, 12)$. Find the probability that they drink the recommended amount, $8 \times 8oz$ on a given day. Use TWO methods: (1) R, and (2) standardizing to a Z score and getting the probability from the table in the back of the book.
5. In one page (one side), tell me about your data, and give me some relevant summary statistics and figures that tell the basic story so far. The figures need to be large enough to demonstrate your point, but small enough so that you still have room to tell me something interesting. Don't bore me with loads and loads statistics and figures – just tell me what's up with your data set in an interesting way.

May 20, 2010

1. **Working with data in R.**

The function **read.table** gives you a data frame. For a data frame named `my.df.name`, elements are accessed like matrix, i.e., `my.df.name[x,y]`... see what leaving either `x` or `y` blank does. The function **names** can be used on a data frame. For a data.frame column named `my.name`, try `my.dataframe$my.name`. The function **length** does different things for different objects. The function **na.omit** can be called on a data frame. Missing values in R should be denoted by `NA` so that R recognizes them as missing. The function **help** will probably be very useful for this problem.

Get 'Vaccine_Data.txt' from my website:

- (a) As is, how many rows does the data set have?
 - (b) What are the values in the first row of this data set?
 - (c) How are missing values denoted in this data set?
Drop the missing values for the remaining questions.
 - (d) How many rows does the data set have now?
 - (e) What is `mean(your.dataframename$age)`?
2. **Get some data** =) Load it into R. *You'll use this data throughout the course, so choose something you like... but you will be able to change your data set later if you need to.*
3. Show that the function **mean** in R calculates \bar{x} for a sample x . Hint: Use the **sum** function to help calculate \bar{x} .
4. Show that the function **var** in R calculates s^2 for a sample x .
5. Give me a histogram and a boxplot of an interesting variable in *your* data.
6. What is the mean, median, mode, variance, and standard deviation of that variable?
7. From the last problem, what is the difference between the mean and median, and why is that?

FOR THE REMAINING QUESTIONS **help(rnorm)** and **help(rbinom)** will be useful.

8. Show using simulation that \bar{x} is an estimator of $E[X]$.
9. Show using simulation that $\sum_{i=1}^n \frac{(x_i - \bar{x})^2}{n-1}$ is an estimator of $\text{Var}[X]$.
10. For a variable X sampled from a binomial model with parameters N and p , the book says $E[X] = Np$ and $\text{Var}[X] = Np(1-p)$. Show this is likely true using your previous two answers.
11. For a binomial model with $N = 10$ and $p = .7$, what is the probability that a single sampled value from the model will be 5 or greater? What is the 80th percentile?
12. For a normal model with $\mu = 10$ and $\sigma^2 = .7$, what is the probability that a single sampled value from the model will be 5 or greater? What is the 20th percentile?
13. For either the discrete or continuous case, prove mathematically that $\text{Var}[X] = E[X^2] - E[X]^2$

May 19, 2010

1. **Would you like to change your mind?**

There are three doors. If you choose the right door, your wildest dreams will come true – if not, you get a goat. Assume for that sake of argument that your wildest dreams do not include making feta cheese. Your guide instructs you to pick a door, so you do. Your guide then opens one of the doors you did not pick, and low and behold, there's a goat. Your guide proceeds to profusely beg you to reconsider your pick, and exhorts you to take the remaining door instead of the door you originally chose. Would you like to change your mind?

Completely comment the following R code to show me you know what every line does. Then use the code to answer the question.

```
n = 10
truth = t(rmultinom(n,1,c(1/3,1/3,1/3)))#help(rmultinom)
truth = truth%%seq(1,3,1)
guess = 1
number_correct_NO_SWITCH = 0
number_correct_SWITCH = 0
for(i in 1:n)
{
  #not swiching strategy wins if
  if(guess == truth[i]) #otherwise guessing loses.
  {
    number_correct_NO_SWITCH = number_correct_NO_SWITCH + 1
  }
  #swiching strategy wins if
  if(guess != truth[i]) #otherwise switching loses
  {
    number_correct_SWITCH = number_correct_SWITCH + 1
  }
}
number_correct_NO_SWITCH/n
number_correct_SWITCH/n
```

2. **Working with data in R.**

```
read.table
# Gives you a data frame. For a data frame named my.df.name
# elements are accessed like matrix: my.df.name[x,y]
# ...See what leaving either x or y blank does.
names
# For a data.frame column named my.name, try my.dataframe$my.name
length
# Does different things for different objects...
na.omit
# Missing values in R are denoted by NA.
```

Get 'Vaccine.Data.txt' from my website:

- (a) How many rows does the data set have?
 - (b) What are the values in the first row of this data set?
 - (c) How are missing values denoted in this data set?
Drop the missing values for the remaining questions.
 - (d) How many rows does the data set have?
 - (e) What is $\text{mean}(\text{your.dataframename}\$age)$?
3. **Get some data** (=) Load it into R. *You'll use this data throughout the course, so choose something you like. You'll be able to change your data set later if you need, but you need SOMETHING by TOMORROW or you won't be able to do the assignment : ... <*

The ideal data set will have continuous, categorical, and binary data.

Ask your favorite professor <http://infochimps.org/> <http://goduke.statsgeek.com/>
Ask a professor you know <http://www.data.gov/> <http://theinfo.org/>

4. **Coin flipping and Independent/Mutually Exclusive events**

- (a) If I flip a fair coin 3 times, what is the probability of getting all heads?
- (b) How about getting heads-tails-heads, in that order?
- (c) What is the probability of seeing two heads in three flips?
- (d) For each question above, what was the key rule used in calculating the probability?

Drawing time. Show this equality with a picture:

$$\Pr(A \text{ or } B \text{ or } C) \stackrel{?}{=} \Pr(A) + \Pr(B) + \Pr(C) - \Pr(AB) - \Pr(BC) - \Pr(AC) + \Pr(ABC).$$

5. **Do we share birthdays?**

Let's assume there are 20 people in class, counting me. What's the probability that at least one pair of us have the same birthday? Assume no one was born on the extra day of a leap year. Hint: $\Pr(A) = 1 - \Pr(\text{not } A)$.

6. **Prove mathematically that** $\Pr(A|B)\Pr(B) + \Pr(A|\text{not } B)\Pr(\text{not } B) = \Pr(A)$.

7. **Bayes theorem.** Suppose there is a disease screening test that is fairly reliable, with

$$\Pr(\text{test} = +|\text{disease}) = \Pr(\text{test} = -|\text{healthy}) = 9/10.$$

Suppose also that disease is quite rare, with $\Pr(\text{disease}) = 1/10000$. What's the probability an individual with a positive test result actually has the disease, i.e., $\Pr(\text{disease}|\text{test} = +)$?