Review for Exam 1

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

Exam 1

- Designed to be 1 hour long. You'll have the entire 80 minutes.
- You may bring one 4 by 6 notecard as described in our email. This will be turned in with your exam and is worth 5 points. (Be sure to write your name on the card.)
- Lots of practice problems posted on the 18,05x site and on the repository.

Topics

1. Sets.
2. Counting.
3. Sample space, outcome, event, probability function.
4. Probability: conditional probability, independence, Bayes theorem.
5. Discrete random variables: events, pmf, cdf.
6. Bernoulli(p), binomial(n, p), geometric(p), uniform(n)
7. $E(X)$, $\text{Var}(X)$, σ
8. Continuous random variables: pdf, cdf.
9. uniform(a,b), exponential(λ), normal(μ,σ)
10. Transforming random variables.
11. Quantiles.
12. Central limit theorem, law of large numbers, histograms.
13. Joint distributions: pmf, pdf, cdf, covariance and correlation.

Sets and counting

- Sets:
  $\emptyset$, union, intersection, complement Venn diagrams, products

- Counting:
  inclusion-exclusion, rule of product,
  permutations ${}_nP_k$, combinations ${}_nC_k = \binom{n}{k}$

Probability

- Sample space, outcome, event, probability function.
  Rule: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.
  Special case: $P(A^c) = 1 - P(A)$
  (A and B disjoint $\Rightarrow P(A \cup B) = P(A) + P(B)$.)

- Conditional probability, multiplication rule, trees, law of total probability, independence

- Bayes' theorem, base rate fallacy

Random variables, expectation and variance

- Discrete random variables: events, pmf, cdf
- Bernoulli(p), binomial(n, p), geometric(p), uniform(n)
- $E(X)$, meaning, algebraic properties, $E(h(X))$
- Var(X), meaning, algebraic properties
- Continuous random variables: pdf, cdf
- uniform(a,b), exponential(λ), normal(μ,σ)
- Transforming random variables
- Quantiles

Central limit theorem

- Law of large numbers averages and histograms
- Central limit theorem

Joint distributions

- Joint pmf, pdf, cdf.
- Marginal pmf, pdf, cdf
- Covariance and correlation.

Problem correlation

1. Flip a coin 3 times. Use a joint pmf table to compute the covariance and correlation between the number of heads on the first 2 and the number of heads on the last 2 flips.

2. Flip a coin 5 times. Use properties of covariance to compute the covariance and correlation between the number of heads on the first 3 and last 3 flips.

<u>**answer:**</u> 1. Let $X =$ the number of heads on the first 2 flips and Y the number in the last 2. Considering all 8 possibe tosses: HHH, HHT etc we get the following joint pmf for X and Y

Y/X	0	1	2	
0	1/8	1/8	0	1/4
1	1/8	1/4	1/8	1/2
2	0	1/8	1/8	1/4
	1/4	1/2	1/4	1

Solution continued on next slide

Solution 1 continued

Using the table we find

$$E(XY) = \frac{1}{4} + 2\frac{1}{8} + 2\frac{1}{8} + 4\frac{1}{8} = \frac{5}{4}.$$

We know $E(X) = 1 = E(Y)$ so

$$\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{5}{4} - 1 = \frac{1}{4}.$$

Since X is the sum of 2 independent Bernoulli(.5) we have $\sigma_X = \sqrt{2/4}$

$$\mathrm{Cor}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(2)/4} = \frac{1}{2}.$$

Solution to 2 on next slide

Solution 2

2. As usual let X_i = the number of heads on the i^{th} flip, i.e. 0 or 1. Let $X = X_1 + X_2 + X_3$ the sum of the first 3 flips and $Y = X_3 + X_4 + X_5$ the sum of the last 3. Using the algebraic properties of covariance we have

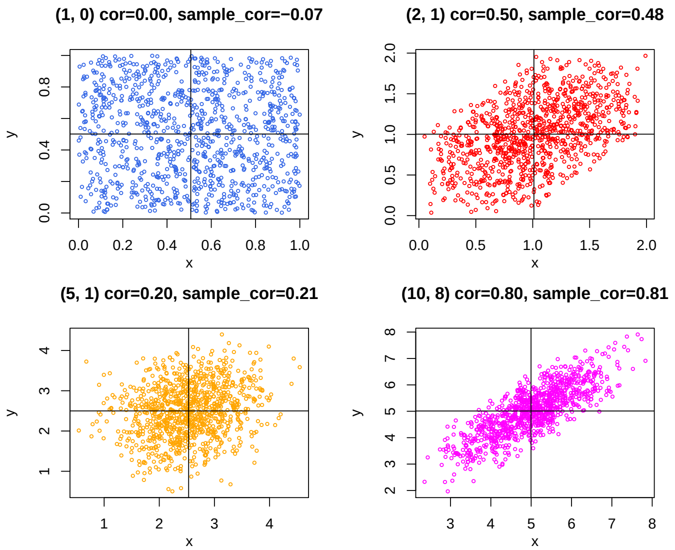
$$\begin{aligned}
\text{Cov}(X, Y) &= \text{Cov}(X_1 + X_2 + X_3, X_3 + X_4 + X_5) \\
&= \text{Cov}(X_1, X_3) + \text{Cov}(X_1, X_4) + \text{Cov}(X_1, X_5) \\
&+ \text{Cov}(X_2, X_3) + \text{Cov}(X_2, X_4) + \text{Cov}(X_2, X_5) \\
&+ \text{Cov}(X_3, X_3) + \text{Cov}(X_3, X_4) + \text{Cov}(X_3, X_5)
\end{aligned}$$

Because the X_i are independent the only non-zero term in the above sum is $\text{Cov}(X_3 X_3) = \text{Var}(X_3) = \dfrac{1}{4}$ Therefore, $\text{Cov}(X, Y) = \frac{1}{4}$.
We get the correlation by dividing by the standard deviations. Since X is the sum of 3 independent Bernoulli(.5) we have $\sigma_X = \sqrt{3/4}$

$$\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(3)/4} = \frac{1}{3}.$$

Sums of overlapping uniform variables

(See class 7 slides for details)



Hospitals (binomial, CLT, etc)

- A certain town is served by two hospitals.
- Larger hospital: about 45 babies born each day.
- Smaller hospital about 15 babies born each day.
- For a period of 1 year, each hospital recorded the days on which more than 60% of the babies born were boys.

(a) Which hospital do you think recorded more such days?
(i) The larger hospital. (ii) The smaller hospital.
(iii) About the same (that is, within 5% of each other).

(b) Let L_i (resp., S_i) be the Bernoulli random variable which takes the value 1 if more than 60% of the babies born in the larger (resp., smaller) hospital on the i^{th} day were boys. Determine the distribution of L_i and of S_i.

Continued on next slide

Hospital continued

(c) Let L (resp., S) be the number of days on which more than 60% of the babies born in the larger (resp., smaller) hospital were boys. What type of distribution do L and S have? Compute the expected value and variance in each case.

(d) Via the CLT, approximate the .84 quantile of L (resp., S). Would you like to revise your answer to part (a)?

(e) What is the correlation of L and S? What is the joint pmf of L and S? Visualize the region corresponding to the event $L > S$. Express $P(L > S)$ as a double sum.

answer: (a) When this question was asked in a study, the number of undergraduates who chose each option was 21, 21, and 55, respectively. This shows a lack of intuition for the relevance of sample size on deviation from the true mean (i.e., variance).

Continued on next slide.

Solution continued

(b) The random variable X_L, giving the number of boys born in the larger hospital on day i, is governed by a Bin$(45, .5)$ distribution. So L_i has a Ber(p_L) distribution with

$$p_L = P(X > 27) = \sum_{k=28}^{45} \left(\begin{array}{c} 45 \\ k \end{array} \right) .5^{45} \approx .068$$

Similarly, the random variable X_S, giving the number of boys born in the smaller hospital on day i, is governed by a Bin$(15, .5)$ distribution. So S_i has a Ber(p_S) distribution with

$$p_S = P(X_S > 9) = \sum_{k=10}^{15} \left(\begin{array}{c} 15 \\ k \end{array} \right) .5^{15} \approx .151$$

We see that p_S is indeed greater than p_L, consistent with (ii).

Continued on next slide.

Solution continued

(c) Note that $L = \sum_{i=1}^{365} L_i$ and $S = \sum_{i=1}^{365} S_i$. So L has a $\text{Bin}(365, p_L)$ distribution and S has a $\text{Bin}(365, p_S)$ distribution. Thus

$$\begin{aligned} E(L) &= 365 p_L \approx 25 \\ E(S) &= 365 p_S \approx 55 \\ \text{Var}(L) &= 365 p_L(1 - p_L) \approx 23 \\ \text{Var}(S) &= 365 p_S(1 - p_S) \approx 47 \end{aligned}$$

(d) mean + sd in each case:
For L, $q_{.84} \approx 25 + \sqrt{23}$.
For S, $q_{.84} \approx 55 + \sqrt{47}$.

Continued on next slide.

Solution continued

(e) Since L and S are independent, their joint distribution is determined by multiplying their individual distributions. Both L and S are binomial with $n = 365$ and p_L and p_S computed above. Thus

$$p_{l,s} P(L = i \text{ and } S = j) = p(i,j) = \binom{365}{i} p_L^i (1 - p_L)^{365-i} \binom{365}{j} p_S^j (1 - p_S)^{365}$$

Thus

$$P(L > S) = \sum_{i=0}^{364} \sum_{j=i+1}^{365} p(i,j) \approx .0000916$$

We used the R code on the next slide to do the computations.

R code

```
pL = 1 - pbinom(.6*45,45,.5)
pS = 1 - pbinom(.6*15,15,.5)
print(pL)
print(pS)

pLGreaterS = 0
for(i in 0:365)
{
    for(j in 0:(i-1))
    {
        = pLGreaterS + dbinom(i,365,pL)*dbinom(j,365,pS)
    }
}
print(pLGreaterS)
```

Counties with high kidney cancer death rates

Map of kidney cancer death rate removed due to copyright restrictions.

Counties with low kidney cancer death rates

Map of kidney cancer death rate removed due to copyright restrictions.

Discussion and reference on next slide

Discussion

The maps were taken from
Teaching Statistics: A Bag of Tricks by Andrew Gelman, Deborah Nolan

- The first map shows with the lowest 10% age-standardized death rates for cancer of kidney/ureter for U.S. white males 1980-1989.
- The second map shows the highest 10%
- We see that both maps are dominated by low population counties. This reflects the higher variability around the national mean rate among low population counties and conversely the low variability about the mean rate among high population counties. As in the hospital example this follows from the central limit theorem.

Lottery (expected value)

Consider the following situations:

(a) Would you accept a gamble that offers a 10% chance to win \$95 and a 90% chance of losing \$5?

(b) Would you pay \$5 to participate in a lottery that offers a 10% percent chance to win \$100 and a 90% chance to win nothing?

What is the expected value of your change in assets in each case?

Discussion on next slide

Discussion: Framing bias / cost versus loss

- The two situations are identical with an expected value of gaining \$5.
- Study: 132 undergrads were given these questions (in different orders) separated by a short filler problem.
- 55 gave different preferences to the two events.
- 42 rejected (a) but accepted (b).

One interpretation is that we are far more willing to pay a cost up front than risk a loss.*

*See *Judgment under uncertainty: heuristics and biases* by Tversky and Kahneman.