# Leftover: Independence, Covariance and Correlation

18.05 Spring 2017

<table>
<thead>
<tr>
<th>$X \backslash Y$</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>2</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>3</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>4</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>5</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
<tr>
<td>6</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
</tr>
</tbody>
</table>
Independence

Events $A$ and $B$ are independent if

$$P(A \cap B) = P(A)P(B).$$

Random variables $X$ and $Y$ are independent if

$$F(x, y) = F_X(x)F_Y(y).$$

Discrete random variables $X$ and $Y$ are independent if

$$p(x_i, y_j) = p_X(x_i)p_Y(y_j).$$

Continuous random variables $X$ and $Y$ are independent if

$$f(x, y) = f_X(x)f_Y(y).$$
Concept question: independence I

Roll two dice: \( X = \) value on first, \( Y = \) value on second

<table>
<thead>
<tr>
<th>( X \setminus Y )</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>( p(x_i) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
<tr>
<td>2</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
<tr>
<td>3</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
<tr>
<td>4</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
<tr>
<td>5</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
<tr>
<td>6</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
</tbody>
</table>

| \( p(y_j) \)         | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1/6 | 1            |

Are \( X \) and \( Y \) independent? 1. Yes 2. No

**answer:** 1. Yes. Every cell probability is the product of the marginal probabilities.
Concept question: independence II

Roll two dice: \( X = \) value on first, \( T = \) sum

<table>
<thead>
<tr>
<th>( X ) ( \backslash T )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/36</td>
<td>1/6</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>( p(x_i) )</th>
<th>1/6</th>
</tr>
</thead>
</table>

|--------------|--------------------------------------------------------|

Are \( X \) and \( Y \) independent?  
1. Yes  
2. No

answer: 2. No. The cells with probability zero are clearly not the product of the marginal probabilities.
Concept Question

Among the following pdf’s which are independent? (Each of the ranges is a rectangle chosen so that \( \int \int f(x, y) \, dx \, dy = 1 \).)

(i) \( f(x, y) = 4x^2y^3 \).
(ii) \( f(x, y) = \frac{1}{2}(x^3y + xy^3) \).
(iii) \( f(x, y) = 6e^{-3x-2y} \)

Put a 1 for independent and a 0 for not-independent.

(a) 111  (b) 110  (c) 101  (d) 100
(e) 011  (f) 010  (g) 001  (h) 000

answer: (c). Explanation on next slide.
Solution

(i) Independent. The variables can be separated: the marginal densities are \( f_X(x) = ax^2 \) and \( f_Y(y) = by^3 \) for some constants \( a \) and \( b \) with \( ab = 4 \).

(ii) Not independent. \( X \) and \( Y \) are not independent because there is no way to factor \( f(x, y) \) into a product \( f_X(x)f_Y(y) \).

(iii) Independent. The variables can be separated: the marginal densities are \( f_X(x) = ae^{-3x} \) and \( f_Y(y) = be^{-2y} \) for some constants \( a \) and \( b \) with \( ab = 6 \).
Covariance

Measures the degree to which two random variables vary together, e.g. height and weight of people.

\( X, Y \) random variables with means \( \mu_X \) and \( \mu_Y \).

\[
\text{Cov}(X, Y) = \text{def} \ E((X - \mu_X)(Y - \mu_Y)).
\]

- Vary together might mean \( X \) is usually bigger than \( \mu_X \) when \( Y \) is bigger than \( \mu_Y \), and vice versa. In this case \((X - \mu_X)(Y - \mu_Y)\) is usually positive, so \( \text{Cov}(X, Y) \) is positive.

- Vary together might mean \( X \) is usually bigger than \( \mu_X \) when \( Y \) is smaller than \( \mu_Y \), and vice versa. In this case \((X - \mu_X)(Y - \mu_Y)\) is usually negative, so \( \text{Cov}(X, Y) \) is negative.

- If \( X \) and \( Y \) don’t vary together, then sign of \((X - \mu_X)\) tells nothing about sign of \((Y - \mu_Y)\). In this case \((X - \mu_X)(Y - \mu_Y)\) can be both positive and negative, so \( \text{Cov}(X, Y) \) might be zero or small.
Properties of covariance

Properties

1. \( \text{Cov}(aX + b, cY + d) = ac\text{Cov}(X, Y) \) for constants \( a, b, c, d \).
2. \( \text{Cov}(X_1 + X_2, Y) = \text{Cov}(X_1, Y) + \text{Cov}(X_2, Y) \).
3. \( \text{Cov}(X, X) = \text{Var}(X) \)
4. \( \text{Cov}(X, Y) = E(XY) - \mu_X \mu_Y \).
5. If \( X \) and \( Y \) are independent then \( \text{Cov}(X, Y) = 0 \).
6. **Warning** The converse is not true: if covariance is 0, the variables might not be independent.
Concept question

Suppose we have the following joint probability table.

<table>
<thead>
<tr>
<th>Y \ X</th>
<th>-1</th>
<th>0</th>
<th>1</th>
<th>(p(y_j))</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>1/2</td>
<td>0</td>
<td>1/2</td>
</tr>
<tr>
<td>1</td>
<td>1/4</td>
<td>0</td>
<td>1/4</td>
<td>1/2</td>
</tr>
</tbody>
</table>

| \(p(x_i)\) | 1/4 | 1/2 | 1/4 | 1   |

At your table work out the covariance \(\text{Cov}(X, Y)\).

Because the covariance is 0 we know that \(X\) and \(Y\) are independent.

1. True 2. False

Key point: covariance measures the linear relationship between \(X\) and \(Y\). It can completely miss a quadratic or higher order relationship.
Board question: computing covariance

Flip a fair coin 12 times.

Let $X = \text{number of heads in the first 7 flips}$

Let $Y = \text{number of heads on the last 7 flips}$.

Compute Cov$(X, Y)$,
Solution

Use the properties of covariance.

\( X_i \) = the number of heads on the \( i^{\text{th}} \) flip. (So \( X_i \sim \text{Bernoulli}(0.5) \).)

\[
X = X_1 + X_2 + \ldots + X_7 \quad \text{and} \quad Y = X_6 + X_7 + \ldots + X_{12}.
\]

We know \( \text{Var}(X_i) = 1/4 \). Therefore using Property 2 (linearity) of covariance

\[
\text{Cov}(X, Y) = \text{Cov}(X_1 + X_2 + \ldots + X_7, X_6 + X_7 + \ldots + X_{12})
\]

\[
= \text{Cov}(X_1, X_6) + \text{Cov}(X_1, X_7) + \text{Cov}(X_1, X_8) + \ldots + \text{Cov}(X_7, X_{12})
\]

Since the different tosses are independent we know

\[
\text{Cov}(X_1, X_6) = 0, \quad \text{Cov}(X_1, X_7) = 0, \quad \text{Cov}(X_1, X_8) = 0, \quad \text{etc.}
\]

Looking at the expression for \( \text{Cov}(X, Y) \) there are only two non-zero terms

\[
\text{Cov}(X, Y) = \text{Cov}(X_6, X_6) + \text{Cov}(X_7, X_7) = \text{Var}(X_6) + \text{Var}(X_7) = \frac{1}{2}.
\]
Correlation

Like covariance, but removes scale.
The correlation coefficient between \( X \) and \( Y \) is defined by

\[
\text{Cor}(X, Y) = \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y}.
\]

1. \( \rho = \) covariance of standardized versions of \( X \) and \( Y \).
2. \( \rho \) is dimensionless (it’s a ratio).
3. \(-1 \leq \rho \leq 1\).
4. \( \rho = 1 \) if and only if \( Y = aX + b \) with \( a > 0 \).
5. \( \rho = -1 \) if and only if \( Y = aX + b \) with \( a < 0 \).
Real-life correlations

- Over time, amount of ice cream consumption is correlated with number of pool drownings.
- In 1685 (and today) being a student is the most dangerous profession.
- In 90% of bar fights ending in a death the person who started the fight died.
- Hormone replacement therapy (HRT) is correlated with a lower rate of coronary heart disease (CHD).

Discussion is on the next slides.
Real-life correlations discussion

- Ice cream does not cause drownings. Both are correlated with summer weather.

- In a study in 1685 of the ages and professions of deceased men, it was found that the profession with the lowest average age of death was “student.” But, being a student does not cause you to die at an early age. Being a student means you are young. This is what makes the average of those that die so low.

- A study of fights in bars in which someone was killed found that, in 90% of the cases, the person who started the fight was the one who died.

  Of course, it’s the person who survived telling the story.

Continued on next slide
In a widely studied example, numerous epidemiological studies showed that women who were taking combined hormone replacement therapy (HRT) also had a lower-than-average incidence of coronary heart disease (CHD), leading doctors to propose that HRT was protective against CHD. But randomized controlled trials showed that HRT caused a small but statistically significant increase in risk of CHD. Re-analysis of the data from the epidemiological studies showed that women undertaking HRT were more likely to be from higher socio-economic groups (ABC1), with better-than-average diet and exercise regimens. The use of HRT and decreased incidence of coronary heart disease were coincident effects of a common cause (i.e. the benefits associated with a higher socioeconomic status), rather than cause and effect, as had been supposed.
Correlation is not causation

Edward Tufte: ”Empirically observed covariation is a necessary but not sufficient condition for causality.”
Overlapping sums of uniform random variables

We made two random variables $X$ and $Y$ from overlapping sums of uniform random variables

For example:

$$X = X_1 + X_2 + X_3 + X_4 + X_5$$
$$Y = X_3 + X_4 + X_5 + X_6 + X_7$$

These are sums of 5 of the $X_i$ with 3 in common.

If we sum $r$ of the $X_i$ with $s$ in common we name it $(r, s)$.

Below are a series of scatterplots produced using R.
Scatter plots

(1, 0) cor=0.00, sample_cor=-0.07

(2, 1) cor=0.50, sample_cor=0.48

(5, 1) cor=0.20, sample_cor=0.21

(10, 8) cor=0.80, sample_cor=0.81
Toss a fair coin $2n + 1$ times. Let $X$ be the number of heads on the first $n + 1$ tosses and $Y$ the number on the last $n + 1$ tosses.

If $n = 1000$ then $\text{Cov}(X, Y)$ is:

(a) 0    (b) 1/4    (c) 1/2    (d) 1

(e) More than 1    (f) tiny but not 0

**answer:** 2. 1/4. This is computed in the answer to the next table question.
Board question

Toss a fair coin $2n + 1$ times. Let $X$ be the number of heads on the first $n + 1$ tosses and $Y$ the number on the last $n + 1$ tosses.

Compute $\text{Cov}(X, Y)$ and $\text{Cor}(X, Y)$.

As usual let $X_i$ = the number of heads on the $i^{\text{th}}$ flip, i.e. 0 or 1. Then

$$X = \sum_{1}^{n+1} X_i, \quad Y = \sum_{n+1}^{2n+1} X_i$$

$X$ is the sum of $n + 1$ independent Bernoulli($1/2$) random variables, so

$$\mu_X = E(X) = \frac{n+1}{2}, \quad \text{and} \quad \text{Var}(X) = \frac{n+1}{4}.$$  

Likewise, $\mu_Y = E(Y) = \frac{n+1}{2}$, and $\text{Var}(Y) = \frac{n+1}{4}$.

*Continued on next slide.*
Solution continued

Now,

\[
\text{Cov}(X, Y) = \text{Cov} \left( \sum_{1}^{n+1} X_i \sum_{n+1}^{2n+1} X_j \right) = \sum_{i=1}^{n+1} \sum_{j=n+1}^{2n+1} \text{Cov}(X_i; X_j).
\]

Because the \(X_i\) are independent the only non-zero term in the above sum is \(\text{Cov}(X_{n+1}X_{n+1}) = \text{Var}(X_{n+1}) = \frac{1}{4}\) Therefore,

\[
\text{Cov}(X, Y) = \frac{1}{4}.
\]

We get the correlation by dividing by the standard deviations.

\[
\text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(n+1)/4} = \frac{1}{n+1}.
\]

This makes sense: as \(n\) increases the correlation should decrease since the contribution of the one flip they have in common becomes less important.
Review for Exam 1
18.05 Spring 2017
Extra office hours

- Tuesday:
  - David 3–5 in 2-355
  - Watch web site for more
- Friday, Saturday, Sunday March 10–12: no office hours
Exam 1

- Designed to be 1 hour long. You’ll have the entire 80 minutes.

- You may bring one 4 by 6 notecard. This will be turned in with your exam. (Be sure to write your name on the card.)

- Lots of practice problems posted on class web site.

- No calculators. (They won’t be necessary.)

- Be sure to get familiar with the table of normal probabilities (it’s easy).
## Normal Table

**Standard normal table of left tail probabilities.**

<table>
<thead>
<tr>
<th>$z$</th>
<th>$\Phi(z)$</th>
<th>$z$</th>
<th>$\Phi(z)$</th>
<th>$z$</th>
<th>$\Phi(z)$</th>
<th>$z$</th>
<th>$\Phi(z)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>-4.00</td>
<td>0.0000</td>
<td>-2.00</td>
<td>0.0228</td>
<td>0.00</td>
<td>0.5000</td>
<td>2.00</td>
<td>0.9772</td>
</tr>
<tr>
<td>-3.95</td>
<td>0.0000</td>
<td>-1.95</td>
<td>0.0256</td>
<td>0.05</td>
<td>0.5199</td>
<td>2.05</td>
<td>0.9798</td>
</tr>
<tr>
<td>-3.90</td>
<td>0.0000</td>
<td>-1.90</td>
<td>0.0287</td>
<td>0.10</td>
<td>0.5398</td>
<td>2.10</td>
<td>0.9821</td>
</tr>
<tr>
<td>-3.85</td>
<td>0.0001</td>
<td>-1.85</td>
<td>0.0322</td>
<td>0.15</td>
<td>0.5596</td>
<td>2.15</td>
<td>0.9842</td>
</tr>
<tr>
<td>-3.80</td>
<td>0.0001</td>
<td>-1.80</td>
<td>0.0359</td>
<td>0.20</td>
<td>0.5793</td>
<td>2.20</td>
<td>0.9861</td>
</tr>
<tr>
<td>-3.75</td>
<td>0.0001</td>
<td>-1.75</td>
<td>0.0401</td>
<td>0.25</td>
<td>0.5987</td>
<td>2.25</td>
<td>0.9878</td>
</tr>
<tr>
<td>-3.70</td>
<td>0.0001</td>
<td>-1.70</td>
<td>0.0446</td>
<td>0.30</td>
<td>0.6179</td>
<td>2.30</td>
<td>0.9893</td>
</tr>
<tr>
<td>-3.65</td>
<td>0.0001</td>
<td>-1.65</td>
<td>0.0495</td>
<td>0.35</td>
<td>0.6368</td>
<td>2.35</td>
<td>0.9906</td>
</tr>
<tr>
<td>-3.60</td>
<td>0.0002</td>
<td>-1.60</td>
<td>0.0548</td>
<td>0.40</td>
<td>0.6554</td>
<td>2.40</td>
<td>0.9918</td>
</tr>
<tr>
<td>-3.55</td>
<td>0.0002</td>
<td>-1.55</td>
<td>0.0606</td>
<td>0.45</td>
<td>0.6736</td>
<td>2.45</td>
<td>0.9929</td>
</tr>
<tr>
<td>-3.50</td>
<td>0.0002</td>
<td>-1.50</td>
<td>0.0666</td>
<td>0.50</td>
<td>0.6915</td>
<td>2.50</td>
<td>0.9938</td>
</tr>
<tr>
<td>-3.45</td>
<td>0.0003</td>
<td>-1.45</td>
<td>0.0735</td>
<td>0.55</td>
<td>0.7088</td>
<td>2.55</td>
<td>0.9946</td>
</tr>
<tr>
<td>-3.40</td>
<td>0.0003</td>
<td>-1.40</td>
<td>0.0808</td>
<td>0.60</td>
<td>0.7257</td>
<td>2.60</td>
<td>0.9953</td>
</tr>
<tr>
<td>-3.35</td>
<td>0.0004</td>
<td>-1.35</td>
<td>0.0885</td>
<td>0.65</td>
<td>0.7422</td>
<td>2.65</td>
<td>0.9960</td>
</tr>
<tr>
<td>-3.30</td>
<td>0.0005</td>
<td>-1.30</td>
<td>0.0968</td>
<td>0.70</td>
<td>0.7580</td>
<td>2.70</td>
<td>0.9965</td>
</tr>
<tr>
<td>-3.25</td>
<td>0.0006</td>
<td>-1.25</td>
<td>0.1056</td>
<td>0.75</td>
<td>0.7734</td>
<td>2.75</td>
<td>0.9970</td>
</tr>
</tbody>
</table>
Today

- David will work examples on one side of the room.
- Gus and Sam and Lucas will hold office hours on the other side of the room.
- You should feel free to go back and forth between the sides.
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. $E(X)$, $\text{Var}(X)$, $\sigma$
9. uniform($a,b$), exponential($\lambda$), normal($\mu,\sigma^2$)
10. Transforming random variables.
11. Quantiles.
12. Central limit theorem, law of large numbers, histograms.
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($\lambda$), normal($\mu,\sigma$)
- 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.
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) Assume exactly 45 and 15 babies are born at the hospitals each day. 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
(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 0.84 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.

Solution on next slide.
Solution

**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).

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

$$p_L = P(X_\cdot > 27) = \sum_{k=28}^{45} \binom{45}{k} .5^{45} \approx 0.068.$$  

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

$$p_S = P(X_S > 9) = \sum_{k=10}^{15} \binom{15}{k} .5^{15} \approx 0.151.$$  

We see that $p_S$ is indeed greater than $p_L$, consistent with (ii).
(c) Note that $L = \sum_{i=1}^{365} L_i$ and $S = \sum_{i=1}^{365} S_i$. So $L$ has a Bin(365, $p_L$) distribution and $S$ has a Bin(365, $p_S$) distribution. Thus

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

(d) By the CLT, the 0.84 quantile is approximately the mean + one sd in each case:
For $L$, $q_{0.84} \approx 25 + \sqrt{23}$.
For $S$, $q_{0.84} \approx 55 + \sqrt{47}$.

*Continued on next slide.*
(e) Since $L$ and $S$ are independent, their correlation is 0 and 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

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

Thus

$$P(L > S) = \sum_{i=0}^{364} \sum_{j=i+1}^{365} p(i, j) \approx 0.0000916$$

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

```r
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
Counties with low kidney cancer death rates

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.
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.

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

$$
\begin{array}{c|c|c|c|c|c|c|c|c}
Y/X & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 \\
\hline
0 & 1/8 & 1/8 & 0 & 1/8 & 1/8 & 0 & 1/8 & 1/8 \\
1 & 1/8 & 1/4 & 1/8 & 1/2 & 1/4 & 1/8 & 1/4 & 1/2 \\
2 & 0 & 1/8 & 1/4 & 1/8 & 1/4 & 1/8 & 1/4 & 1/2 \\
3 & 1/4 & 1/2 & 1/4 & 1 & 1/2 & 1/4 & 1/8 & 1/4 \\
\end{array}
$$

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

\[ \text{Cov}(X, Y) = E(XY) - E(X)E(Y) = \frac{5}{4} - 1 = \frac{1}{4}. \]

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

\[ \text{Cor}(X, Y) = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{1/4}{(2)/4} = \frac{1}{2}. \]

Solution to 2 on next slide
Solution 2

2. As usual let \( X_i \) = the number of heads on the \( i^{\text{th}} \) 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

\[
\text{Cov}(X, Y) = \text{Cov}(X_1 + X_2 + X_3, X_3 + X_4 + X_5) \\
= \text{Cov}(X_1, X_3) + \text{Cov}(X_1, X_4) + \text{Cov}(X_1, X_5) \\
+ \text{Cov}(X_2, X_3) + \text{Cov}(X_2, X_4) + \text{Cov}(X_2, X_5) \\
+ \text{Cov}(X_3, X_3) + \text{Cov}(X_3, X_4) + \text{Cov}(X_3, X_5)
\]

Because the \( X_i \) are independent the only non-zero term in the above sum is \( \text{Cov}(X_3, X_3) = \text{Var}(X_3) = \frac{1}{4} \) 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 \( \sigma_X = \sqrt{3/4} \)

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