Political Science Math Camp:
Problem Set 4

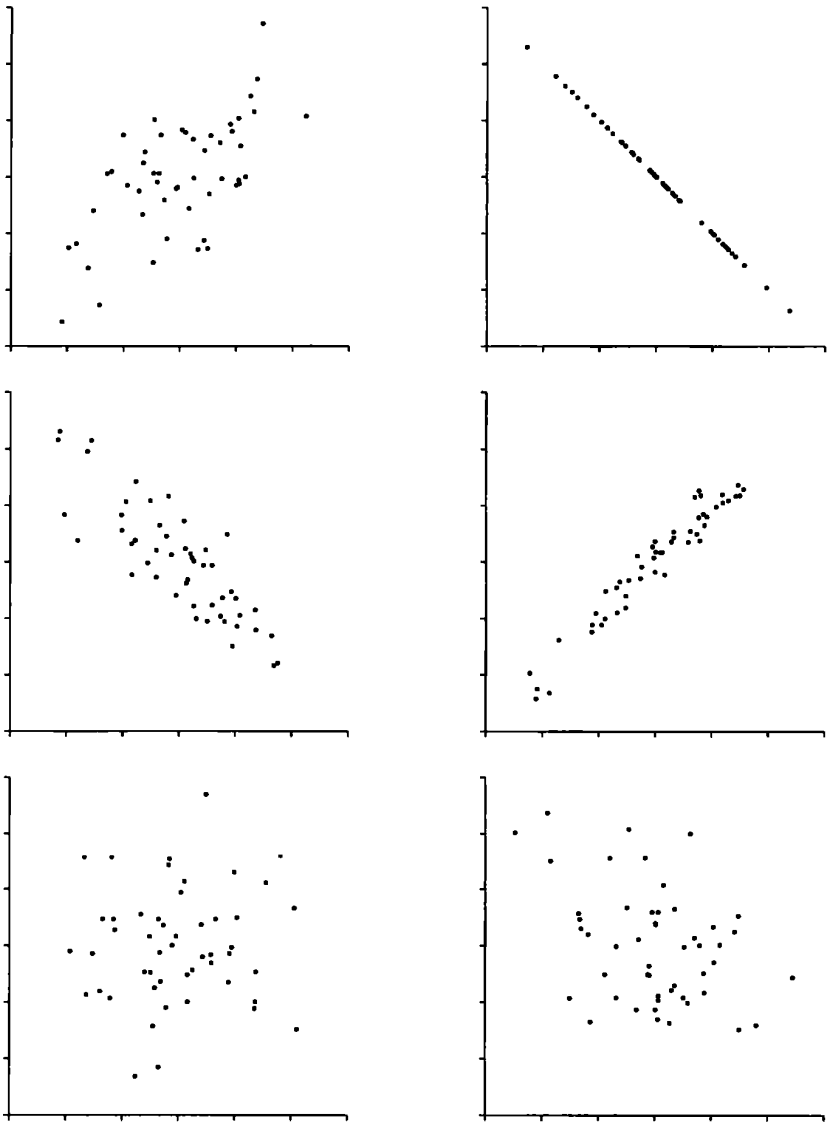
Due on Monday, Aug 14th at 9:00 am

1. Find the length of the following vectors.
   (a) (3,4)
   (b) (0,-3)
   (c) (1,1,1)
   (d) (-1,-1)
   (e) (1, 2, 3)
   (f) (3, 0, 0, 0)
2. Find the length of $\mathbf{u} = (1, 2, -1)$ and $\mathbf{v} = (1, 2, 4)$. Are there vectors orthogonal?
3. If $\mathbf{u}$ and $\mathbf{v}$ are $n \times 1$, show that $||u + v||^2 = ||u||^2 + ||v||^2 + 2u \cdot v$.
4. If u and v are $nx1$, show that $||u+v||^2 = ||u||^2 + ||v||^2$ if and only if $u \perp v$. (This is Pythagoras' theorem in n dimensions.)
5. Let $\mathbf{u} = (1, 2)$, $\mathbf{v} = (0, 1)$, $\mathbf{w} = (1, -3)$, $\mathbf{x} = (1, 2, 0)$, and $z = (0, 1, 1)$. Compute the following vectors whenever they are defined:
   (a) $\mathbf{u} + \mathbf{v}$
   (b) $-4\mathbf{w}$
   (c) $\mathbf{u} + \mathbf{z}$
   (d) $3\mathbf{z}$
   (e) $2\mathbf{v}$
   (f) $\mathbf{u} + 2\mathbf{v}$
   (g) $\mathbf{u} - \mathbf{v}$
   (h) $3\mathbf{x} + \mathbf{z}$
6. Suppose that X_1 and X_2 are independent variables with mean 0 and variance σ^2.
   (a) What is the covariance of $X_1 + 2X_2$ and $4X_1 - 3X_2$?
   (b) What is the correlation?
7. In the following, assume X_1, X_2, X_3, X_4 are independent random variables with $Cov(X_i, X_j) = \rho\sigma^2$ for $i \neq j$ and σ^2 for $i = j$.
   - What is $Cov(X_1 + X_2, X_1 - X_2)$

- What is $Cov(X_1 + X_3, X_1 - 2X_2 + 4X_3)$? What about the correlation?
- What is $\rho(X_1, X_1 + X_2)$?
- What is $Cov(X_1 - 3X_4, X_2 + 16X_3)$? What is $\rho(X_1 - 3X_4, X_2 + 16X_3)$?

8. For Berkeley freshmen, the average GPA is around 3.0; the SD is about 0.5. The histogram follows the normal curve. Estimate the 30th percentile of the GPA distribution.

9. The figure below has six scatter diagrams for hypothetical data. The correlation coefficients in scrambled order are: $\{-.85, -.38, -1, .06, .97, .62\}$. Match the scaterr diagrams with the correlation coefficient.



10. Two different investigators are working on a growth study. The first measures the heights of 100 children, in inches. The second prefers the metric system, and changes the results to centimeters (multiplying by the conversion factor 2.54 centimeters per inch). A scatter diagram is plotted, showing for each child its height in inches on the horizontal axis and the height in centimeters on the vertical axis.

    - If no mistakes are made in the convertion, what is the correlation?
    - What happens to ρ is mistakes are made in the arithmetic?
    - What happens to ρ if the second investigator goes out and measures the same children again, using metric equipment?

11. Match the lists with the SDs. Explain your reasoning.

(a) $1, -2, -2$	(i) $\sqrt{1/3 \times 2/3}$
(b) $15, 15, 16$	(ii) $2 \times \sqrt{1/3 \times 2/3}$
(c) $-1, -1, -1, 1$	(iii) $3 \times \sqrt{1/3 \times 2/3}$
(d) $0, 0, 0, 1$	(iv) $\times \sqrt{1/4 \times 3/4}$
(e) $0, 0, 2$	(v) $2 \times \sqrt{1/4 \times 3/4}$

12. One hundred draws are made at random with replacement from the box $\{1, 2, 3, 4, 5, 6\}$.

    - If the sum of the draws is 321, what is the average?
    - If the average of the draws is 3.78, what is the sum?
    - Estimate the chance that the average of the draws is between 3 and 4.

13. A group of 50,000 tax forms has an average gross income of \$37,000, with an SD of \$20,000. Furthermore, 20% of the forms have a gross income over \$50,000. A group of 900 forms is chosen at random for audit. To estimate the chance that between 19% and 21% of the forms chosen for audit have gross incomes over \$50,000, a box model is needed.

    - Should the number of tickets in the box be 900 or 50,000?
    - Each ticket in the box shows:
        - A zero or a one
        - A gross income
    - True or false: the SD of the box is \$20,000
    - True or false: the number of draws is 900
    - Find the chance (approximately) that between 19% and 21% of the forms chosen for audit have gross incomes over \$50,000

14. In a certain town, there are 30,000 registered voters, of whom 12,000 are Democrats. A survey organization is about to take a simple random sample of 1,000 registered voters. There is about a 50-50 chance that the percentage of Democrats in the sample will be bigger than _____. Fill in the blank, and explain.

15. A box contains a large number of red and blue marbles, but the proportions are unknown; 100 marbles are drawn at random, and 53 turn out to be red. Say whether each of the following statements is true or false, and explain briefly.

    - The percentage of red marbles in the box can be estimated as 53%; the SE is 5%.

- The 5% measures the likely size of the chance error in the 53%.
- The 53% is likely to be off the percentage of red marbles in the box, by 5% or so.
- A 95%-confidence interval for the percentage of red marbles in the box is 43% to 63%.
- A 95%-confidence interval for the percentage of red marbles in the sample is 43% to 63%.

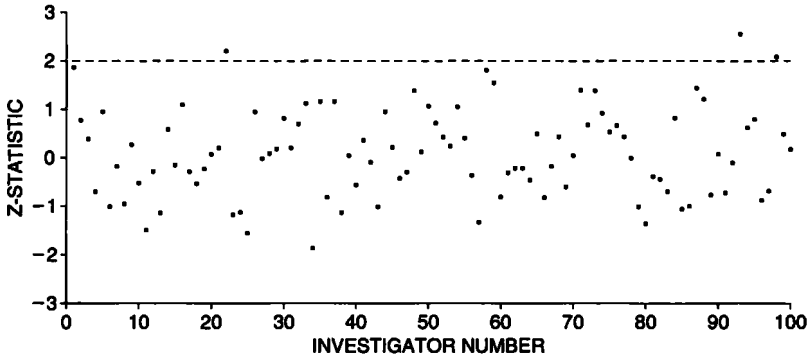
16. A simple random sample of 1,000 persons is taken to estimate the percentage of Democrats in a large population. It turns out that 543 of the people in the sample are Democrats. True or false, and explain:

    - The sample percentage is (543/1,000) x 100% = 54.3%; the SE for the sample percentage is 1.6%.
    - 54.3% ± 3.2% is a 95%-confidence interval for the population percentage.
    - 54.3% ± 3.2% is a 95%-confidence interval for the sample percentage.
    - There is about a 95% chance for the percentage of Democrats in the population to be in the range 54.3% ± 3.2%.

17. A real estate office wants to make a survey in a certain town, which has 50,000 households, to determine how far the head of household has to commute to work. A simple random sample of 1,000 households is chosen, the occupants are interviewed, and it is found that on average, the heads of the sample households commuted 8.7 miles to work; the SD of the distances was 9.0 miles. (All distances are one-way; if someone isn't working, the commute distance is defined to be 0.)

    - The average commute distance of all 50,000 heads of households in the town is estimated as _____, and this estimate is likely to be off by _____ or so.
    - If possible, find a 95%-confidence interval for the average commute distance of all heads of households in the town. If this isn't possible, explain why not.

18. One hundred investigators set out to test the null hypothesis that the average of the numbers in a certain box equals 50. Each investigator takes 250 tickets at random with replacement, computes the average of the draws, and does a z-test. The results are plotted in the diagram. Investigator #1 got a z-statistic of 1.9, which is plotted as the point (1, 1.9). Investigator #2 got a z-statistic of 0.8, which is plotted as (2, 0.8), and so forth. Unknown to the investigators, the null hypothesis is true.



(a) True or false, and explain: the z-statistic is positive when the average of the draws is more than 50.

(b) How many investigators should get a positive z-statistic?

(c) How many of them should get a z-statistic bigger than 2? How many of them actually do?

(d) If $z = 2$, what is P?

19. A coin is tossed 10,000 times, and it lands heads 5,167 times. Is the chance of heads equal to 50%? Or are there too many heads for that?

- Formulate the null and alternative hypotheses in terms of a box model.
- Compute z and P.
- What do you conclude?

20. The Gallup poll asks respondents how they would rate the honesty and ethical standards of people in different fields—very high, high, average, low, or very low. The percentage who rated clergy "very high or high" dropped from 60% in 2000 to 54% in 2005. This may have been due to scandals involving sex abuse; or it may have been a chance variation. (You may assume that in each year, the results are based on independent simple random samples of 1,000 persons in each year.)

- Should you make a one-sample z-test or a two-sample z-test? Why?
- Formulate the null and alternative hypotheses in terms of a box model. Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?
- Do you need one box or two? Why? How many tickets go into each box? How many draws? What do the tickets show? What do the null and alternative hypotheses say about the box(es)?
- Can the difference between 60% and 54% be explained as a chance variation? Or was it the scandals? Or something else?

21. One hundred draws are made at random with replacement from box A, and 250 are made at random with replacement from box B.

- 50 of the draws from box A are positive, compared to 131 from box B: 50.0% versus 52.4%. Is this difference real, or due to chance?
- The draws from box A average 1.4 and their SD is 15.3; the draws from box B average 6.3 and their SD is 16.1. Is this difference between the averages statistically significant?