Political Science Math Camp:
Problem Set 4

Answer Key

1. **Find the length of the following vectors.**

   (a) **(3,4)**

   $$||(3,4)||= \sqrt{9+16} = 5$$

   (b) **(0,-3)**

   $$||(0,-3)||= \sqrt{9} = 3$$

   (c) **(1,1,1)**

   $$||(1,1,1)||= \sqrt{1+1+1} = \sqrt{3}$$

   (d) **(-1,-1)**

   $$||(-1,-1)||= \sqrt{1+1} = \sqrt{2}$$

   (e) **(1, 2, 3)**

   $$||(1,2,3)||= \sqrt{1+4+9} = \sqrt{14}$$

   (f) **(3, 0, 0, 0)**

   $$||(3,0,0,0)||= \sqrt{9} = 3$$

2. **Find the length of $\mathbf{u} = (1,2,-1)$ and $\mathbf{v} = (1,2,4)$. Are there vectors orthogonal?**
   We first need the legnth of both vectors:

   $$\begin{aligned} ||(1,2,-1)|| &= \sqrt{1+4+1} \\ &= \sqrt{6} \\ ||(1,2,4)|| &= \sqrt{1+4+16} \\ &= \sqrt{21} \end{aligned}$$

   Two column vectors u, v are orthogonal when $u'v = 0$.

$$\begin{aligned} u'v &= 1 \times 1 + 2 \times 2 - 1 \times 4 \\ &= 1 + 4 - 4 \\ &= 1 \end{aligned}$$

Thus, u and v are not orthogonal vectors.

3. **If u and v are $n \times 1$, show that $||u+v||^2 = ||u||^2 + ||v||^2 + 2u \cdot v$.**

   The norm of a vector is denoted $||u||$, with $||u||^2$ equal to the inner product $u \cdot u$. Thus,

$$\begin{aligned} ||u||^2 &= u \cdot u \\ &= u'u \\ &= (u_1^2 + u_2^2 + ... + u_n^2), \end{aligned}$$

   and therefore

$$\begin{aligned} ||u+v||^2 &= (u+v) \cdot (u+v) \\ &= (u+v)'(u+v) \\ &= u'u + v'v + 2u'v \\ &= ||u||^2 + ||v||^2 + 2u \cdot v \end{aligned}$$

   For the third line, note that $u'v = v'u = \sum_{i=1}^n u_i v_i$.

   Another way to see this is as follows. Note that the $n \times 1$ vector $(u+v)$ has typical element $(u_i + v_i)$, so the inner product $(u+v) \cdot (u+v)$ has typical element $(u_i + v_i)^2$. In fact,

$$\begin{aligned} (u+v) \cdot (u+v) &= \sum_{i=1}^n (u_i + v_i)^2 \\ &= \sum_{i=1}^n (u_i^2 + 2u_i v_i + v_i^2) \\ &= \sum_{i=1}^n u_i^2 + 2\sum_{i=1}^n (u_i v_i) + \sum_{i=1}^n v_i^2) \\ &= \|u\|^2 + 2u \cdot v + \|v\|^2 \end{aligned} \tag{1}$$

   The second line follows from writing out terms, while the third follows from distributing the summation sign and the fourth follows from the definition of the inner product.

4. **If u and v are $n \times 1$, show that $||u+v||^2 = ||u||^2 + ||v||^2$ if and only if $u \perp v$.**

   The "if" direction follows from the final line of the previous exercise: if $u \perp v$, $2u \cdot v = 0$, so $(u+v) \cdot (u+v) = \|u\|^2 + \|v\|^2$. The "only if" direction is similar: by (1), $\|u+v\|^2 = \|u\|^2 + \|v\|^2 + 2u \cdot v$, and $2u \cdot v$ is equal to zero only if $u \cdot v = 0$. But that's the definition of orthogonality, so $\|u+v\|^2 = \|u\|^2 + \|v\|^2$ only if $u \perp v$.

   Comment: 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) $\mathbf{u} + \mathbf{v} = (1,3)$

(b) $-4\mathbf{w} =$ (-4, 12)

(c) $\mathbf{u} + \mathbf{z}$. Not defined, the vectors have different length.

(d) $3\mathbf{z} = (0,3,3)$

(e) $2\mathbf{v} = (0,2)$

(f) $\mathbf{u} + 2\mathbf{v} = (1,4)$

(g) $\mathbf{u} - \mathbf{v} = (1,1)$

(h) $3\mathbf{x} + \mathbf{z} = (3,7,1)$

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 $4X_1 - 3X_2$?**
Recall $Cov(X + Y, Z) = Cov(X, Z) + Cov(Y, Z)$ so

$$\begin{aligned} Cov(X_1 + 2X_2, 4X_1 - 3X_2) &= Cov(X_1, 4X_1 - 3X_2) + Cov(2X_2, 4X_1 - 3X_2) \\ &= Cov(X_1, 4X_1) + Cov(X_1, -3X_2) + Cov(2X_2, 4X_1) + \\ &\quad Cov(2X_2, -3X_2) \end{aligned}$$

And $Cov(cX, Y) = c \times Cov(X, Y)$ thus,

$$\begin{aligned} Cov(X_1 + 2X_2, 4X_1 - 3X_2) &= 4 \times Cov(X_1, X_1) - 3 \times Cov(X_1, X_2) + \\ &\quad 2 \times 4 \times Cov(X_2, X_1) + 2 \times (-3) \times Cov(X_2, X_2) \end{aligned}$$

Since $Cov(X_1, X_1) = Cov(X_2, X_2) = \sigma^2$ and $Cov(X_1, X_2) = 0$,

$$\begin{aligned} Cov(X_1 + 2X_2, 4X_1 - 3X_2) &= 4 \times \sigma^2 - 3 \times 0 + 8 \times 0 - 6 \times \sigma^2 \\ &= 4 \times \sigma^2 - 6 \times \sigma^2 \\ &= -2 \times \sigma^2 \end{aligned}$$

(b) **What is the correlation?**

$$\rho(X, Y) = \frac{Cov(X, Y)}{SD_X \times SD_Y}$$

We have the covariance from above, so we just need the standard deviations. Since X_1, X_2 are independent,

$$\begin{aligned} Var(X_1 + 2X_2) &= Var(X_1) + 2^2 Var(X_2) \\ Var(4X_1 - 3X_2) &= 4^2 Var(X_1) + 3^2 Var(X_2) \end{aligned}$$

And then

$$\begin{aligned} SD(X_1 + 2X_2) &= \sqrt{Var(X_1) + 2^2 Var(X_2)} \\ &= \sqrt{\sigma^2 + 4\sigma^2} \\ &= \sqrt{5}\sigma \\ SD(4X_1 - 3X_2) &= \sqrt{4^2 Var(X_1) + 3^2 Var(X_2)} \\ &= \sqrt{16\sigma^2 + 9\sigma^2} \\ &= \sqrt{25\sigma^2} \\ &= 5\sigma \end{aligned}$$

Putting all together,

$$\begin{aligned} \rho(X, Y) &= \frac{-2 \times \sigma^2}{\sqrt{5}\sigma \times 5\sigma} \\ &= \frac{-2}{\sqrt{5} \times 5} \end{aligned}$$

7. **In the following, assume X_1, X_2, X_3, X_4 are independent random variables with $Cov(X_i, X_j) = \delta\sigma^2$ for $i \neq j$ and σ^2 for $i = j$.**

   This is a tricky problem because it relies on you realizing that since the variables are independence, their covariance, $Cov(X_i, X_j) = \delta\sigma^2$, has to be equal to zero ($\delta = 0$). With this insight, solving each question becomes much easier.

   (a) **What is $Cov(X_1 + X_2, X_1 - X_2)$**

$$\begin{aligned} Cov(X_1 + X_2, X_1 - X_2) &= Cov(X_1, X_1) - Cov(X_1, X_2) + Cov(X_2, X_1) - Cov(X_2, X_2) \\ &= Var(X_1) - Var(X_2) \\ &= 0 \end{aligned}$$

   (b) **What is $Cov(X_1 + X_3, X_1 - 2X_2 + 4X_3)$? What about the correlation?**
   We can ignore the covariance terms for $Cov(X_i, X_j)$ when $i \neq j$:

$$\begin{aligned} Cov(X_1 + X_3, X_1 - 2X_2 + 4X_3) &= Cov(X_1, X_1) + Cov(X_3, 4X_3) \\ &= Var(X_1) + 4 \times Var(X_3) \\ &= \sigma^2 + 4 \times \sigma^2 \\ &= 5 \times \sigma^2 \end{aligned}$$

   And the correlation...

$$\begin{aligned}
\rho(X_1 + X_3, X_1 - 2X_2 + 4X_3) &= \frac{Cov(X_1 + X_3, X_1 - 2X_2 + 4X_3)}{SD(X_1 + X_3) \times SD(X_1 - 2X_2 + 4X_3)} \\
SD(X_1 + X_3) &= \sqrt{Var(X_1) + Var(X_3)} \\
&= \sqrt{2\sigma^2} \\
SD(X_1 - 2X_2 + 4X_3) &= \sqrt{Var(X_1) + 4Var(X_2) + 16Var(X_3)} \\
&= \sqrt{21\sigma^2} \\
\rho(X_1 + X_3, X_1 - 2X_2 + 4X_3) &= \frac{5\sigma^2}{\sqrt{2\sigma^2} \times \sqrt{21\sigma^2}} \\
&= \frac{5}{\sqrt{2} \times \sqrt{21}}
\end{aligned}$$

(c) **What is $\rho(X_1, X_1 + X_2)$?**

$$\rho(X_1, X_1 + X_2) = \frac{Cov(X_1, X_1 + X_2)}{SD(X_1) \times SD(X_1 + X_2)}$$

We first need to find $Cov(X_1, X_1 + X_2)$:

$$\begin{aligned}
Cov(X_1, X_1 + X_2) &= Cov(X_1, X_1) + Cov(X_1, X_2) \\
&= Var(X_1) + Cov(X_1, X_2) \\
&= \sigma^2 + 0 \\
&= \sigma^2
\end{aligned}$$

$$\begin{aligned}
SD(X_1) &= \sqrt{Var(X_1)} \\
&= \sqrt{\sigma^2} \\
&= \sigma
\end{aligned}$$

$$\begin{aligned}
SD(X_1 + X_2) &= \sqrt{Var(X_1) + Var(X_2) + 2 \times Cov(X_1, X_2)} \\
&= \sqrt{\sigma^2 + \sigma^2 + 0} \\
&= \sqrt{2} \times \sigma
\end{aligned}$$

$$\begin{aligned}\rho(X_1, X_1 + X_2) &= \frac{\sigma^2}{\sigma \times \sqrt{2}\sigma} \\ &= \frac{\sigma^2}{\sqrt{2}\sigma^2} \\ &= \frac{1}{\sqrt{2}}\end{aligned}$$

(d) **What is $Cov(X_1 - 3X_4, X_2 + 16X_3)$? What is $\rho(X_1 - 3X_4, X_2 + 16X_3)$?**
Since the covariances when $i \neq j$ are zero, we just get:

$$Cov(X_1 - 3X_4, X_2 + 16X_3) = 0$$

And since the covariance is the numerator for the correlation, the correlation will also be zero.

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.**

We are looking for the GPA such that

$$\Phi\Big[\frac{GPA - 3.0}{0.5}\Big] = Pr\Big[X \leq \frac{GPA - 3.0}{0.5}\Big] = .3$$

We could look in the standard normal table for the Z value associated to .3 or we can simply ask `R` using the `qnorm()` function.

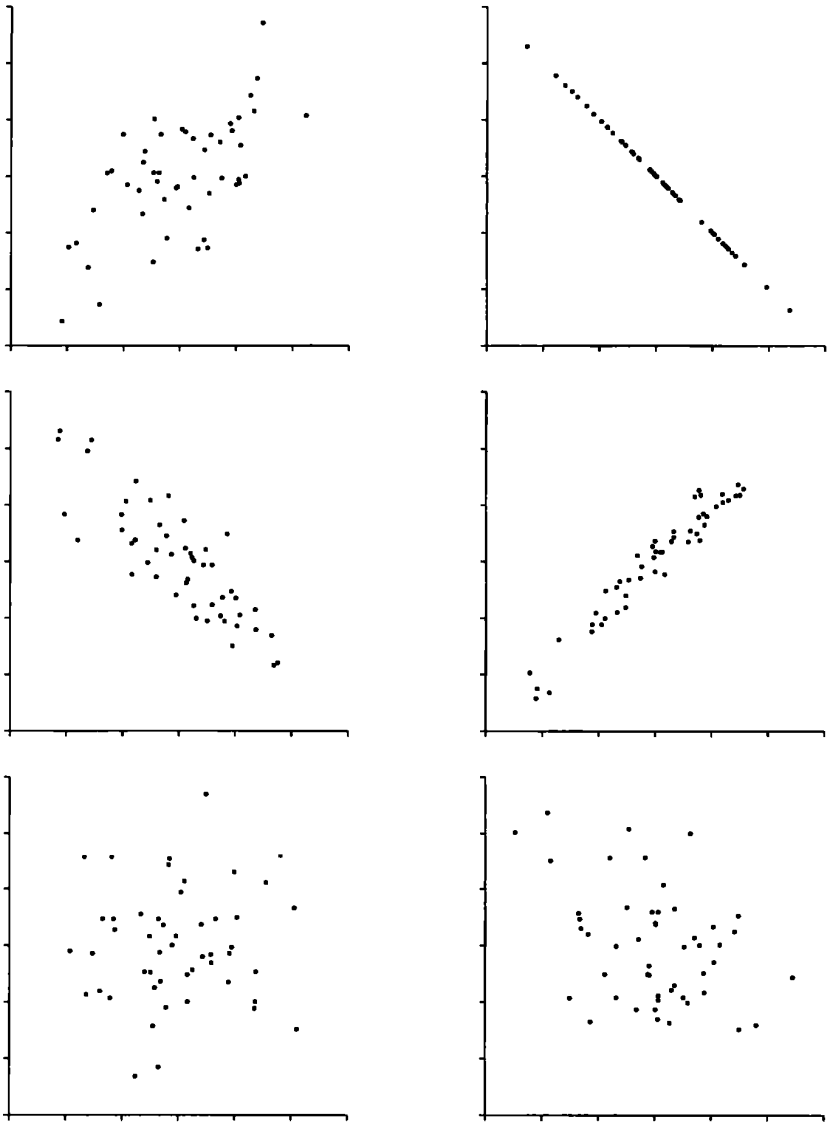
```
qnorm(.3)

## [1] -0.5244005
```

Since $\Phi[Z = -.524] = .3$, we can just solve for GPA:

$$\begin{aligned}\frac{GPA - 3.0}{0.5} &= -.524 \\ GPA &= -.524 \times 0.5 + 3 \\ &= 2.738\end{aligned}$$

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 scatter diagrams with the correlation coefficient.**

From top to bottom and left to right, the correlation coefficients associated with each point are $\{.62, -1, -.85, .97, .06, -.38\}$



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 conversion, what is the correlation?**
      If no mistakes are made, the correlation will be one as they just express the same measure in difference units.
    - **What happens to ρ is mistakes are made in the arithmetic?**
      A mistake in the arithmetic would reduce the correlation coefficient (in absolute value).
    - **What happens to ρ if the second investigator goes out and measures the same children again, using metric equipment?**
      This would also reduce the correlation coefficient in absolute value as the new measuring device could introduce measurement error.

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) $\sqrt{1/4 \times 3/4}$
(e) $0, 0, 2$	(v) $2 \times \sqrt{1/4 \times 3/4}$

    Here we can use the rule in FPP for boxes with two types of tickets.

$$(\text{Large number} - \text{Small number})\sqrt{\text{Prop. large number} \times \text{Prop. small number}}$$

    So we have:

    - $(a) \to (iii)$
    - $(b) \to (i)$
    - $(c) \to (v)$
    - $(d) \to (iv)$
    - $(e) \to (ii)$

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

    (a) **If the sum of the draws is 321, what is the average?**
    The average will just be $\frac{321}{100} = 3.21$.

    (b) **If the average of the draws is 3.78, what is the sum?**
    The sum will be $\overline{X} \times 100 = 378$

    (c) **Estimate the chance that the average of the draws is between 3 and 4.**
    First, we need to calculate the population standard deviation.

$$SD = \sqrt{\frac{(1-3.5)^2 + (2-3.5)^2 + (3-3.5)^2 + (4-3.5)^2 + (5-3.5)^2 + (6-3.5)^2}{6}}$$
$$= 1.708$$

Now we need to find the standard error.

$$SE = \frac{SD}{\sqrt{n}}$$
$$= \frac{1.871}{\sqrt{100}}$$
$$= 0.1708$$

Now, we need to find a z-score for 3 and a z-score for 4.
$z(3) = \frac{3-3.5}{0.1708} = -2.927$ $z(4) = \frac{4-3.5}{0.1708} = 2.927$
Now we find the area under the normal distribution curve to the left of the critical value $z = 2.927$ and subtract off the area to the left of $z = -2.927$.

```
    pnorm(2.927)
```

```
## [1] 0.9982888
```

```
pnorm(-2.927)
```

```
## [1] 0.001711244
```

```
pnorm(2.927)-pnorm(-2.927)
```

```
## [1] 0.9965775
```

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 are 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.**

    (a) **Should the number of tickets in the box be 900 or 50,000?**
    The box should have the complete population of tax forms, so there should be 50,000 tickets in the box.

    (b) **Each ticket in the box shows a zero or a one? Or gross income?**
    Each ticket shows a zero or a one. Tickets with gross income over $50,000 show a one, while the rest shows a zero.

    (c) **True or false: the SD of the box is $20,000.**
    False. The SD of the 50,000 tax forms is $20,000, but the tickets in the box don't have the gross income but zeros and ones—depending on whether gross income is over $50,000. Since 20% of the forms have gross income over $50,000, the probability of drawing a ticket with a one is $p = .2$. Thus, the SD of the box is
    $$\sqrt{p(1-p)} = \sqrt{.2 \times (1-.2)} = .4.$$
    . To get the SD of the percentage, we need to multiply this by 100, so the SD is 40%.

(d) **True or false: the number of draws is 900.**

True, 900 tickets are drawn from the box without replacement.

(e) **Find the chance (approximately) that between 19% and 21% of the forms chosen for audit have gross incomes over $50,000.**

The problem is asking for the area under the curve between 19 and 21%. However, to find the area we need to standardized this values. To be able to do this, we are missing the SE of the sample mean. From (c), we know the mean of the box $V[x] = .4^2$. To get the SE of the sample mean, we need

$$\begin{aligned} SE(\bar{X}) &= \sqrt{\hat{\bar{X}}} \\ &= \sqrt{\frac{V[X]}{n}} \\ &= \sqrt{\frac{.4^2}{900}} \\ &= 0.013 \end{aligned}$$

We now have

$$\begin{aligned} Pr(.19 \leq \bar{X} \leq .21) &= \Phi\Big[\frac{.21 - .2}{0.013}\Big] - \Phi\Big[\frac{.19 - .2}{0.013}\Big] \\ &= \Phi\Big[\frac{.01}{0.013}\Big] - \Phi\Big[\frac{-.01}{0.013}\Big] \\ &= \Phi\Big[0.769\Big] - \Phi\Big[-0.769\Big] \end{aligned}$$

```
pnorm(0.769) - pnorm(-0.769)

## [1] 0.5581067
```

There is approximately a 56% chance 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 50-50 chance that the percentage of Democrats in the sample will be bigger than _____. Fill in the blank, and explain.**

    There will be a 50-50 chance that the percentage of Democrats in the sample is bigger that the expected value of the sample percentage. Since there are 12,000 Democrats among the 30,000 registered voters, that means that 40% of the registered voters are Democrats. The expected value of the percentage of the sample is equal to the percentage of the population and will thus also be 40%.

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

(a) **The percentage of red marbles in the box can be estimates as 53%; the SE is 5%.**

This is true. The percentage of marbles in the sample is an unbiased estimator for the percentage of marbles in the box.

The sample proportion of red marbles is also an estimator for p, the probability of drawing a red marble from the box. Thus, we can estimate the variance of the box as $\hat{V}[X] = \hat{p}(a - \hat{p}) = .53(1 - .53) = 0.249$. The SE for $\hat{p}$ is then $\sqrt{.249} = .49$. In percentages, this is about 5%.

(b) **The 5% measures the likely size of the chance error in 53%.**

True.

(c) **The 53% is likely to be off the percentage of red marbles in the box by 5% or so.**

True. This is about a 68% confidence interval, following the normal distribution.

(d) **A 95% confidence interval for the percentage of red marbles in the box is 43% to 63%.**

True. Since the sampling distribution of the population percentage follows the normal distribution, a 95% confidence interval for the population paramenter can be built considering about 2 (1.96) standard deviations away from the estimated parameter. In this case 2 standard deviations means about 10%.

(e) **A 95% confidence interval for the percentage of red marbles in the sample is 43% to 63%.**

This is false. We know with certainty that the percentage of red marbles in the sample is 53%, we don't need a confidence interval for that.

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.**

(a) **The sample percentage is** $(543/1,000) \times 100\% = 54.3\%$ **the SE for the sample percentage is** 1.6%**.**

True. This follows the same logic as 13 (a).

(b) **54.3% ± 3.2% is a 95% confidence interval for the population percentage.**

True. See 13 (d).

(c) **54.3% ± 3.2% is a 95% confidence interval for the sample percentage.**

False. The sample percentage is known with certainty.

(d) **There is about 95% chance for the percentage of Democrats in the population to be in the range 54.3% ± 3.2%.**

False. If you view the interval as fixed, the chance is either 0 or 1. The chances are in the sampling procedure, not the population. That is why statisticians use the term "confidence interval".

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 household commuted 8.7 miles to work; the SD of the distances was 9 miles.**

(a) **What is the estimated average commute distance of all 50,000 households in the town? By how much is this estimate likely to be off?**

The estimated average commute distance is 8.7.

The SE of the average commute is

$$SE(\bar{X}) = \sqrt{\frac{9^2}{1000}} = \frac{9}{\sqrt{1000}} = .28$$

(b) **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.**

A 95% confidence interval is $(8.15, 9.25)$.

```
8.7 + 1.96 * .28
```

```
## [1] 9.2488
```

```
8.7 - 1.96 * .28
```

```
## [1] 8.1512
```

18. **One hundred investigators set out to test the null hypothesis that the average of the number 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. Investigar #1 got a z-statistic of 1.9 which is plotted at the point $(1, 1.9)$. Investigator #2 got a z-statistic of 0.8, which is plotted as $(2, .8)$, 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.**

True. $z = \frac{observed - expected}{SE}$. Observed is the average of the draws; expected is the average of the box. Since we know that the average of the box is 50, the z-statistic will be positive whenever the average of the draws is more than 50.

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

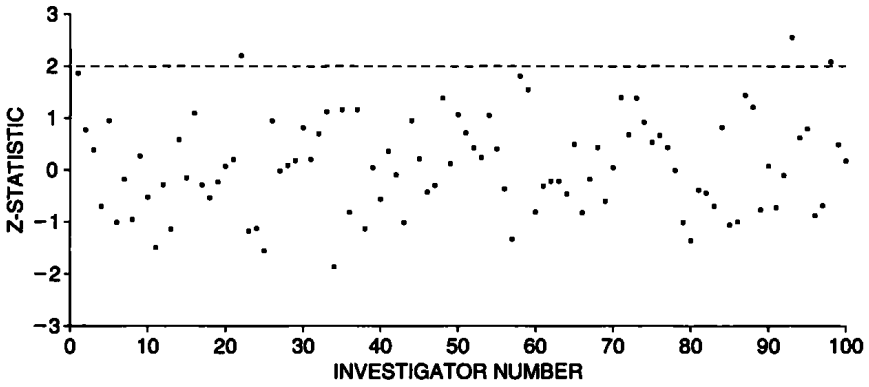
The z-statistic is standardized and centered at 0, so about 50% should get a positive z-statistic. With 100 investigators, that should be about 50.

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

About 2% of the investigators should get a z-statistic bigger than 2 and 3 of them do.

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

If z is 2, P will be about 2.5% (one tailed).


Z-STATISTIC
3
2
1
0
−1
−2
−3
0
10
20
30
40
50
60
70
80
90
100
INVESTIGATOR NUMBER

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?**

    (a) **Formulate the null and the alternative hypotheses in terms of a box model.**
    Here, we are basically testing whether or not the coin being flipped is fair. The null hypothesis can be represented with a box with two "tickets"—one for heads and the other for tails. We are drawing 10,000 tickets at random with replacement. The null hypothesis is that the probability of drawing each ticket is equal to .5. The alternative states that the coin is biased, and the probability of drawing heads is not 0.5. For purposes of being conservative, however, we will formulate this as a two-sided hypothesis test and simply make the alternative hypothesis that the coin is biased.

    (b) **Compute z and P.**
    First, we need to obtain the standard error of the mean of 10,000 coin flips.

$$\begin{aligned}\sigma &= \sqrt{\frac{p(1-p)}{N}} \\ &= \sqrt{\frac{0.5(1-0.5)}{10000}} \\ &= 0.005\end{aligned}$$

    We find the probability of obtaining a heads using the test coin by calculating: $\frac{5167}{10000}$. Thus, $\bar{X} = 0.5167$.

    Now, we can use our standard z-score:

$$\begin{aligned}z &= \frac{0.5167 - 0.50}{0.005} \\ z &= 3.34\end{aligned}$$

    Two tailed p-value:

```
(1 - pnorm(3.34)) *2
```

```
## [1] 0.0008377839
```

    Using a z-score table and a two-sided p-value, P=0.0008, which means there is less than a 0.1% chance that we would observe a value this extreme if the coin were actually fair.

    (c) **What do you conclude?**
    Given that the p-value is far less than the conventional level of significance required to reject the null (i.e. 0.05), we confidently reject the null that the coin is fair. Therefore, we can fairly safely say that we observe too many heads for the coin to be fair.

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 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.**

(a) **Should you make a one-sample z-test or a two sample z-test?**

There are two independent random samples, so the model will have two boxes. Thus, a two sample z-test is the appropriate test.

(b) **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)?**

As stated in the previous question, the model will have two boxes, one for 2000 and the other for 2005. Each box has as many tickets as the population for each respective year. The boxes have 0-1 tickets, where 1 means that the respondent rated clergy "very high or high". We sample 1,000 tickets at random from each box. The null hypothesis is that the proportion of tickets with a 1 in each box is the same, and thus their difference is zero. The alternative hypothesis is that the proportion of 1 tickets differs by box (or that it is smaller in the 2005 box).

(c) **Can the difference between 60% and 54% be explained as chance variation? Or was it the scandals? Or something else?**

Here we need to set up a z-statistic to test whether the difference is explained by chance variation:

$$
\begin{aligned}
z &= \frac{(\overline{A} - \overline{B}) - 0}{\sqrt{Var(\overline{A} - \overline{B})}} \\
&= \frac{(.6 - .54)}{\sqrt{Var(\overline{A}) + Var(\overline{B})}} \\
&= \frac{.06}{\sqrt{\frac{Var(A)}{n_A} + \frac{Var(B)}{n_B}}} \\
&= \frac{.06}{\sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{1000} + \frac{\hat{p}_B(1-\hat{p}_B)}{1000}}} \\
&= \frac{.06}{\sqrt{\frac{.6(1-.6)}{1000} + \frac{.54(1-.54)}{1000}}} \\
&= \frac{.06}{\sqrt{\frac{.24}{1000} + \frac{.2484}{1000}}} \\
&= \frac{.06}{.022} \\
&= 2.72
\end{aligned}
$$

```
1- pnorm(2.72)
```

```
## [1] 0.003264096
```

```
2*(1- pnorm(2.72))
```

```
## [1] 0.006528192
```

The difference cannot be explained by chance variation. However, we cannot say that the difference is due to the scandals, as other things might have occured between 2000 and 2005 that changed people's perceptions.

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

    (a) **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?**

    Here, each box has two tickets, 0-1, where one is defined as a positive draw. The probability of drawing a one from box A is p_A, the probability of drawing a one from box B is p_B. These are unknown parameters.

    The null hypothesis is that $p_A = p_B$, and that the observed difference is due to chance. We can find a z-score to test this claim.

$$
\begin{aligned}
z &= \frac{(.5 - .524) - 0}{\sqrt{Var(\hat{p_A} - \hat{p_B})}} \\
&= \frac{-.024}{\sqrt{\frac{Var(A)}{n_A} + \frac{Var(B)}{n_B}}} \\
&= \frac{-.024}{\sqrt{\frac{\hat{p_A}(1-\hat{p_A})}{100} + \frac{\hat{p_B}(1-\hat{p_B})}{250}}} \\
&= \frac{-.024}{\sqrt{\frac{\hat{.5}(.5)}{100} + \frac{.524(.476)}{250}}} \\
&= \frac{-.024}{\sqrt{\frac{.25}{100} + \frac{.2494}{250}}} \\
&= \frac{-.024}{\sqrt{0.00349}} \\
&= \frac{-.024}{\sqrt{0.00349}} \\
&= -0.406
\end{aligned}
$$

To get the p-value for this z-stat, we can go to the table for the normal distribution:

```
pnorm(-.406)

## [1] 0.3423713
```

For a two tailed test we have:

```
pnorm(-.406) *2

## [1] 0.6847426
```

There is a 68% probability of observing a difference like the this one under the null hypothesis. Thus, we fail to reject the null that there is no difference in the proportion of positive tickets in each box.

(b) **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 the difference between the averages statistically significant?**

To be able to test this hypothesis we need

$$\begin{aligned} z &= \frac{\text{Observed difference} - \text{Expected difference}}{SE(\text{Difference})} \\ &= \frac{(\overline{A} - \overline{B}) - 0}{SE(\overline{A} - \overline{B})} \\ &= \frac{1.4 - 6.3}{\sqrt{Var(\overline{A} - \overline{B})}} \\ &= \frac{-4.9}{\sqrt{Var(\overline{A}) + Var(\overline{B})}} \\ &= \frac{-4.9}{\sqrt{\frac{Var(A)}{N_A} + \frac{Var(B)}{N_B}}} \\ &= \frac{-4.9}{\sqrt{\frac{Var(A)}{100} + \frac{Var(B)}{250}}} \end{aligned}$$

Where $Var(A), Var(B)$ are unknown parameters (the variance of the box). However, we can use the sample to estimate the variance of the box. Note that to correctly estimate this parameters we should use the formula for the estimation of the variance of the box, with $n-1$ in the denominator. For this problem, it will make little difference, because the n of both samples is large enough that n is very close to $n-1$. But we can do this by taking the variance of the sample, multiplying it by the size of the sample and dividing by $n-1$.

$$\begin{aligned} \hat{Var}(A) &= \frac{n \times SD(sampleA)^2}{n-1} = \frac{100 \times 15.3^2}{99} = 236.45 \\ \hat{Var}(B) &= \frac{n \times SD(sampleB)^2}{n-1} = \frac{250 \times 16.1^2}{249} = 260.25 \end{aligned}$$

So we have:

$$\begin{aligned} z &= \frac{-4.9}{\sqrt{\frac{236.45}{100} + \frac{260.25}{250}}} \\ &= \frac{-4.9}{\sqrt{2.36 + 1.04}} \\ &= \frac{-4.9}{\sqrt{3.4}} \\ &= \frac{-4.9}{1.84} \\ &= -2.66 \end{aligned}$$

We can check with R the cumulative probability associated to the lower tail of the distribution:

```
pnorm(-2.66)

## [1] 0.003907033

pnorm(-2.66)*2

## [1] 0.007814065
```

The two tailed p-value is 0.0078, so the difference is statistically significant. We reject the null hypothesis that there is no difference between the averages.