February 26, 2014
Quiz 4

Your name: Solutions

Quiz rules:

(a) This quiz is closed book, but you are allowed a two-sided sheet of paper of notes and a calculator.
(b) Each question is worth 6 points.
(c) A normal table is provided on the last page.
(d) You have 50 minutes to complete this quiz.
(e) If you fail to show your work and/or explain how you arrived at your answer then no points will be awarded.
(f) You do not need to solve all the problems to do well. Try your best.

1. The Olympic committee is concerned about the rate of banned equipment use. To assess the rate at which banned equipment is used the committee conducts a simple random sample of 400 out of the 50,000 registered Olympic athletes. They asked "Have you knowingly used banned equipment during competition in the last 365 days?"

(a) True or false, and explain briefly: The sample size (which is less than 1% of the total population) is too small to produce an estimate of the percentage with any reasonable confidence of athletes in the population who had knowingly used banned equipment in the last 365 days.

False. In fact, because we have a simple random sample of the population, we can calculate a SE that is quite small. A small SE means we have a high level of precision. It doesn't matter at all that our sample size is just 1% of the total population.

In a percentage problem like this SE = $\frac{\sqrt{p(1-p)}}{\sqrt{\#\ of\ draws}} = \frac{\sqrt{p(1-p)}}{\sqrt{400}}$. We don't know p, but the largest the numerator could be is 0.50, which occurs when p=0.50.

From the above observation we see that the largest the SE could be with a sample size of 400 is:

$$SE = \frac{\sqrt{p(1-p)}}{\sqrt{400}} \geq \frac{\sqrt{0.5(1-0.5)}}{\sqrt{400}} = \frac{0.5}{\sqrt{400}} = 0.025$$

Thus the SE cannot be any larger than 2.5%.

(b) Noting that the statute of limitations for using banned equipment is limited to the prior 365 days (i.e., an athlete cannot be held responsible for actions taken more than 365 days ago), a statistician recommends that the survey's question be changed to: "Excluding competitions in the last 365 days, have you ever knowingly used banned equipment during competition?" Explain how this might reduce at least one kind of bias.

Solution 1: This may reduce **non-response bias**. Athletes may refuse to answer a question which would admit wrong doing and could very well get them barred from the Olympics. By changing the question so the survey respondent would not be admitting to an actionable offense, we may decreasing the probability the respondent will refuse to answer the question.

Solution 2: This may reduce **response bias**. It's possible that, instead of not answering, an athlete may choose to submit a lie – perhaps saying he/she did not engage in illegal activity when in fact he/she had. This would be to protect him/herself from possible repercussions. By changing the question, the athletes have less of a reason to lie because their illegal activities are no longer actionable.

This rewording may be helpful, but it's doubtful that it will do a lot to improve response quality. The suggestion in part c is much better.

(c) The statistician also recommends implementing a system where each athlete rolls a die in private before answering the question. If a 1 or 2 come up then they must answer "no," if a 3 or 4 come up they must answer "yes," and if a 5 or 6 come up they must answer truthfully. The interviewer does not see the outcome of the die. There are two types of bias that this technique has the potential to reduce (i) name the two types of bias (ii) uses 2-3 sentences to say how this technique may reduce these types of bias.

(i) Non-response bias and response bias (a.k.a. "deceptive answers")
(ii) This is a randomized-response survey design. This design reduces these kinds of bias by providing respondents with plausible deniability if their answer is potentially stigmatizing. Here, the athletes who is found to have answered "yes" can reasonably argue that they did so due to the randomness, not because they are admitting to guilt.

2. On January 9, 2014 in West Virginia it was discovered that 4-methylcyclohexane methanol (MCHM) was discovered leaking from a storage tank owned by Freedom Industries into the Elk River and from there into the water supply for Charleston, WV. To get an estimate of the contamination of the water in the area, 10 locations (e.g., streams, drinking wells, ground water) were randomly selected and tested each day for a week. Thus there were 70 tests. Across the tests there was an average level of contamination of 1.1ppb and a standard deviation of 0.2ppb. A local news organization wants to discuss the amount of error in the estimate of the average contamination level so they calculate a standard error of

$$\frac{0.2 \text{ ppb}}{\sqrt{70}} \cong 0.0239 \text{ ppb}$$

Explain the assumptions in the above calculation and take a stance on whether they are justified.

The assumption required for this standard error is that the sample average was generated using either a simple random sample or a measurement error model with Gaussian errors. Neither of these is true here. **Therefore this SE is not justified**.

The 10 locations are randomly sampled, but the 70 data points come from repeated **measures through time**. Measurements across days means there are likely dependencies in our measurements (e.g., the pollution amount is likely changing through time as the pollutant diffuses).

This is also not a measurement error problem with Gaussian errors because of the **spatial dependencies**. One way to think about this is that the 10 locations likely have different amounts of pollution because of their distance from the spill and this means there is no exact value being estimated.

3. It is recommended that people over age 70 have their bone density measured to assess risk for osteoporosis. A common method of measurement is a DXA machine. DXA stands for dual energy x-ray absorptiometry. DXA machines are notoriously error prone, but the Gauss model applies. The chance error has an SD of about 0.5.

   An osteoporosis study is conducted using a simple random sample of 225 people over 70 living in California.

(a) Jerry Attric participates in the survey. The researchers take 16 measurements of his hip using the DXA machine. Jerry's average measurement is 2.7. Can you attach a margin of error to this estimate? If so, do it; otherwise, explain why not.

Yes. The SD of the box is 0.5 with an expected value of 0 (we know that because we're told the Gauss model applies).

We're looking to estimate the margin of error for an average so first we need the SE.

$$SE\ of\ average = \frac{SD\ of\ box}{\sqrt{\#\ of\ draws}} = \frac{0.5}{\sqrt{16}} = 0.125$$

For a margin of error for a 95% CI, we multiply the SE by 2 (or you could be more precise and say 1.96).

$$95\%\ margin\ of\ error = 2 * 0.125 = 0.250$$

(b) The researchers take 16 measurements of the hips of each of the 225 subjects for a total of 3600 measurements. The average measurement is 2.5 across these 3600 measurements. Can you attach a margin of error to this estimate? If so, calculate it; otherwise, explain why not.

No. This is not a measurement error problem because there is **no exact value** that is fixed across all of the sampled 70 year olds. And we cannot use our SE (and therefore our margin of error) that comes from sampling error because there are **two sources of chance** now – (i) measurement error from the machine and (ii) sampling error (which comes from the differences in bone density from person to person).

4. To assess the level of statistical sophistication in the adult American population a simple random sample of 200 American adults was carried out to see who had taken at least one statistics course. A confidence interval of (25.3%, 34.7%) is reported. What confidence level was used?

This is a confidence interval for percentages. Thus it is of the form: $sample\ percentage\ \pm Z_\alpha * SE$. The sample percentage is thus in the middle of the confidence interval. We can find it by adding the upper and lower bounds and dividing by two:

$$sample\ percentage = \frac{0.253 + 0.347}{2} = 0.30$$

We can thus estimate the standard deviation of the box by using the bootstrap.

$$SE\ of\ percentage = \frac{\sqrt{0.30(1-0.30)}}{\sqrt{200}} \cong 0.0324$$

Notice the confidence interval is 0.30 ± 0.047 therefore our margin of error is 0.047.

We can find what confidence level was used by solving for Z_α in the formula

$$margin\ of\ error = Z_\alpha * SE$$

$$0.047 = Z_\alpha * 0.0324$$

$$1.45 = Z_\alpha$$

Which corresponds to a **confidence level of 85.29%** (or just an 85%).

5. You want to know whether the George Washington Bridge scandal has affected Americans' perception of Chris Christie as a candidate. You ask a simple random sample of 1000 Americans the question, "Do you view Chris Christie less favorably as a result of the recent scandal?"

(a) <u>True or false, and explain briefly</u>: there is about a 95% chance that the sample percentage of people who view Chris Christie less favorably will be within 2 SE of the population percentage.

True. This is true because of the Central Limit Theorem and the square root law of sampling. In fact, this is exactly why we're able to build confidence intervals – that sample averages will behave normally with a known SE. This statement is true because it isn't referencing a particular data set, rather **it's making statements about how a sample average behaves in general**. See part b for contrast.

(b) Suppose 235 of the 1000 people surveyed answer "yes" in this particular sample. True or false, and explain briefly: : there is about a 95% chance that 23.5% is within 2.7% of the percentage of all Americans who view Chris Christie less favorably.

False. Looking at your data, constructing a particular confidence interval and then stating that the true percentage has some probability of falling in that particular CI does not make sense. There is no randomness at that point. The true percentage is either in that CI or it is not.

Think of it this way: Before rolling a die, we can talk about the probability of rolling a 3. After rolling a die we can no longer talk (sensibly) about the probability of rolling a 3 on that roll. It either came up as a 3 or it did not. Once we've sampled the data (or performed repeated measurements) all of the randomness is gone and our CI will either cover the true percentage or it won't.

6. A small town with 4,000 households plans to ban household pets in public parks. Rolf Barkley, a dog owner, would like to make a case against this ordinance by citing the average number of pets per household. A recent census of all 4,000 households was taken, but Rolf does not remember the average; he only remembers the SD, which was 2.0. Because he does not have time to survey all 4,000 households, he decides to take a simple random sample of 200 households, in which there was an average of 1.4 pets per household with an SD of 2.5.

(a) Rolf would like to attach an uncertainty to this estimate of 1.4. Is this possible? If so, do it; otherwise, explain why not.

Yes. We have a simple random sample of a population, **without replacement**, so we know how to get an SE for the average. Uncertainty in an estimate is quantified by the SE.

Use the SD of the box, rather than the SD of the sample, because that's how we calculate the SE of average. (Note: when we don't have the SD of the box, then we're forced to use the SD of the sample as a stand-in. We call this approach "bootstrapping.")

$$SE\ of\ average = \frac{SD\ of\ box}{\sqrt{\#of\ draws}} * correction\ factor$$

$$correction\ factor = \frac{\sqrt{4{,}000 - 200}}{\sqrt{4{,}000 - 1}} \cong 0.9748$$

Therefore

$$SE\ of\ average \cong \frac{2.0}{\sqrt{200}} * 0.9748 \cong \mathbf{0.138}$$

Note: Solutions without correction factor are also accepted, in this case SE=0.141.

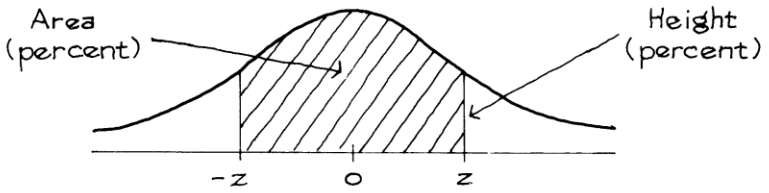
(b) Is it possible to construct a 90% confidence interval for this estimate? If so, do it; otherwise, explain why not.

Yes, it is possible. For a 90% CI we use $Z_\alpha = 1.65$. (See the Z-table for that.)

90% CI is thus: $sample\ ave\ \pm Z_\alpha * SE\ of\ AVE \cong 1.4 \pm 1.65 * 0.138 \cong 1.4 \pm 0.228$.

We could also write it as (1.17, 1.63).

Note: Doesn't change based on correction factor.



A NORMAL TABLE

z	*Height*	*Area*	*z*	*Height*	*Area*	*z*	*Height*	*Area*
0.00	39.89	0	1.50	12.95	86.64	3.00	0.443	99.730
0.05	39.84	3.99	1.55	12.00	87.89	3.05	0.381	99.771
0.10	39.69	7.97	1.60	11.09	89.04	3.10	0.327	99.806
0.15	39.45	11.92	1.65	10.23	90.11	3.15	0.279	99.837
0.20	39.10	15.85	1.70	9.40	91.09	3.20	0.238	99.863
0.25	38.67	19.74	1.75	8.63	91.99	3.25	0.203	99.885
0.30	38.14	23.58	1.80	7.90	92.81	3.30	0.172	99.903
0.35	37.52	27.37	1.85	7.21	93.57	3.35	0.146	99.919
0.40	36.83	31.08	1.90	6.56	94.26	3.40	0.123	99.933
0.45	36.05	34.73	1.95	5.96	94.88	3.45	0.104	99.944
0.50	35.21	38.29	2.00	5.40	95.45	3.50	0.087	99.953
0.55	34.29	41.77	2.05	4.88	95.96	3.55	0.073	99.961
0.60	33.32	45.15	2.10	4.40	96.43	3.60	0.061	99.968
0.65	32.30	48.43	2.15	3.96	96.84	3.65	0.051	99.974
0.70	31.23	51.61	2.20	3.55	97.22	3.70	0.042	99.978
0.75	30.11	54.67	2.25	3.17	97.56	3.75	0.035	99.982
0.80	28.97	57.63	2.30	2.83	97.86	3.80	0.029	99.986
0.85	27.80	60.47	2.35	2.52	98.12	3.85	0.024	99.988
0.90	26.61	63.19	2.40	2.24	98.36	3.90	0.020	99.990
0.95	25.41	65.79	2.45	1.98	98.57	3.95	0.016	99.992
1.00	24.20	68.27	2.50	1.75	98.76	4.00	0.013	99.9937
1.05	22.99	70.63	2.55	1.54	98.92	4.05	0.011	99.9949
1.10	21.79	72.87	2.60	1.36	99.07	4.10	0.009	99.9959
1.15	20.59	74.99	2.65	1.19	99.20	4.15	0.007	99.9967
1.20	19.42	76.99	2.70	1.04	99.31	4.20	0.006	99.9973
1.25	18.26	78.87	2.75	0.91	99.40	4.25	0.005	99.9979
1.30	17.14	80.64	2.80	0.79	99.49	4.30	0.004	99.9983
1.35	16.04	82.30	2.85	9.69	99.56	4.35	0.003	99.9986
1.40	14.97	83.85	2.90	0.60	99.63	4.40	0.002	99.9989
1.45	13.94	85.29	2.95	0.51	99.68	4.45	0.002	99.9991