

Statistics 403 Problem Set 3

Due in lab on Friday, October 2nd

1. Suppose we are working with the sample standard deviation

$$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2},$$

and we initially obtain a data set of 10 values X_1, \dots, X_{10} . We calculate the average value of these 10 numbers to be 6 and the sample standard deviation to be 2. We then add three more values X_{11}, X_{12}, X_{13} to the data set, all of which are equal to 6. What is the sample standard deviation of the new data set?

Solution: From last week's problem set we know that the mean will still be 6 after adding the new data points. We can calculate the variance of the new data points as follows:

$$\begin{aligned} \frac{1}{12} \sum_{i=1}^{13} (X_i - 6)^2 &= \frac{1}{12} \left(\sum_{i=1}^{10} (X_i - 6)^2 + \sum_{i=11}^{13} (X_i - 6)^2 \right) \\ &= \frac{1}{12} \sum_{i=1}^{10} (X_i - 6)^2 \\ &= \frac{9}{12} \left(\frac{1}{9} \sum_{i=1}^{10} (X_i - 6)^2 \right) \\ &= 3. \end{aligned}$$

Thus the sample standard deviation of the new data set is $\sqrt{3}$. Note that this is less than 2. By adding additional points at the mean, the standard deviation always becomes smaller.

2. Suppose we observe 20 iid values from a population with mean 3 and standard deviation 1.5. We average these values to produce an estimate \bar{X} of the population mean EX .
 - (a) What is the probability that our estimate will fall more than 0.5 units from the true value of the population mean?

Solution:

$$\begin{aligned} P(|\bar{X} - 3| > 0.5) &= P(\bar{X} - 3 > 0.5) + P(\bar{X} - 3 < -0.5) \\ &= 2P(\bar{X} - 3 > 0.5) \end{aligned}$$

$$\begin{aligned}
&= 2P\left(\frac{\bar{X} - 3}{1.5/\sqrt{20}} > \frac{0.5}{1.5/\sqrt{20}}\right) \\
&= 2P(Z > 1.49) \\
&\approx 0.14.
\end{aligned}$$

- (b) Suppose that instead of 20 values, we are able to collect as few or as many values as needed so that the probability of \bar{X} being more than 0.5 units from the population mean is 10%. How many values should we collect?

Solution: Let n be the sample size that we will use.

$$\begin{aligned}
P(|\bar{X} - 3| > 0.5) &= P(\bar{X} - 3 > 0.5) + P(\bar{X} - 3 < -0.5) \\
&= 2P(\bar{X} - 3 > 0.5) \\
&= 2P\left(\frac{\bar{X} - 3}{1.5/\sqrt{n}} > \frac{0.5}{1.5/\sqrt{n}}\right) \\
&= 2P\left(Z > \frac{0.5}{1.5/\sqrt{n}}\right) \\
&= 0.1.
\end{aligned}$$

So,

$$P\left(Z > \frac{0.5}{1.5/\sqrt{n}}\right) = 0.05.$$

From a normal distribution table, or using `qnorm(0.95)` in R, we get that

$$P(X > 1.64) = 0.05.$$

Thus we set

$$\frac{0.5}{1.5/\sqrt{n}} = \sqrt{n}/3 = 1.64,$$

and we get $n \approx 24$.

- (c) Repeat part (b) for the situation where the probability of \bar{X} being more than 0.1 units from the population mean must be 10%.

Solution: We get

$$P\left(Z > \frac{0.1}{1.5/\sqrt{n}}\right) = 0.05.$$

So we set

$$\frac{0.1}{1.5/\sqrt{n}} = \sqrt{n}/15 = 1.64,$$

and we get $n \approx 605$.

3. Suppose we are doing a two-sample comparison using data X_1, \dots, X_{10} and Y_1, \dots, Y_{10} sampled independently from two (possibly distinct) populations. Let $D = \bar{X} - \bar{Y}$ be the estimate of the difference in means between the two populations. The standard deviations of the two populations are known to be $\sigma_X = 1$ and $\sigma_Y = 0.7$.

- (a) We have the option of adding five subjects to either the X sample or the Y sample. Which should we choose if the goal is to minimize the variance of D ?

Solution: If we add subjects to the X sample the new variance is

$$1/15 + 0.7^2/10 \approx 0.12.$$

If we add subjects to the Y sample the new variance is

$$1/10 + 0.7^2/15 \approx 0.13.$$

Thus we should add the new data to the X population. This makes sense because the X population is more variable, so in the original data with equal numbers of samples from the X and Y populations, we had less information about the X population than the Y population.

- (b) Now suppose we can add 10 additional subjects to the data set, and we are free to add to the X and Y samples as we like (e.g. we could add 4 subjects to the X sample and 6 subjects to the Y sample). How should we do this so as to minimize the variance of D ? Hint: let $10 + f$ be the new X sample size and let $20 - f$ be the new Y sample size. Treat f as a continuous variable (i.e. ignore the fact that it must be an integer). Then use calculus to minimize $\text{var}(D)$ with respect to f , constraining f to lie between 0 and 10.

Solution:

The function we are minimizing is

$$\text{var}(D) = 1/(10 + f) + 0.7^2/(20 - f).$$

The derivative with respect to f is

$$-1/(10 + f)^2 + 0.7^2/(20 - f)^2.$$

If we set this equal to zero we get

$$0.51f^2 - 49.8f + 351.$$

Solving for f (using the quadratic formula) yields

$$f = \frac{49.8 \pm \sqrt{1764}}{1.02}.$$

The solutions are $f \approx 7.6, 90$. Since we must have $0 \leq f \leq 10$ we need only to check that $f \approx 7.6$ is a local minimum.

Rounding 7.6 to 8, we see that the variance is minimized when we use an X sample size of 18 and a Y sample size of 12.

4. Suppose two research groups are trying to measure an important quantity μ . The first group has an instrument with measurement standard deviation $\sigma = 0.3$ and has funding to collect 10 data points. The second research group has an instrument with measurement standard deviation $\sigma = 0.5$. How many data points should the second research group collect so that its 95% confidence interval for μ is approximately half as wide as the 95% CI from the first group?

Solution: The CI for the first group will be around $4 \cdot 0.3/\sqrt{10} \approx 0.38$ units wide, so the second group needs to have a CI that is 0.19 units wide. The second group's CI will be $4 \cdot 0.5/\sqrt{n}$ units wide, based on a sample of size n . Solving $2/\sqrt{n} = 0.19$ for n yields $n \approx 111$.