Problem Set 5

Conceptual Questions: Predictive and Confidence Intervals

1. True or False: Consider a random variable X. To construct a 95% predictive interval for X, all we need to know is the sampling distribution of X.

2. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% predictive interval for the sampling mean $\bar{X} = \sum_{i=1}^{n} X_i$, all we need to know is the sampling distribution of X.

3. We take a random sample of n independent individuals, and record their outcomes, $X_1, X_2, ..., X_n$. To construct a 95% confidence interval for the true mean of X, denoted μ, all we need to know is the Central Limit Theorem.

Central Limit Theorem and Confidence Intervals According to the WHO Global Database on Anaemia, the mean hemoglobin levels among primary school children in Delhi were estimated at $\mu = 108$ g/L, with standard deviation $\sigma = 12.5$ g/L.

`Source: http://who.int/vmnis/anaemia/data/database/countries/ind_ida.pdf`

Suppose we took a random sample of 75 primary school children in Delhi. Denote the mean hemoglobin levels in this sample as $\bar{x}$. Throughout this question, assume that the sample size is large enough that the central limit theorem is applicable and that σ is known.

1. What is the expected value (mean) of $\bar{x}$?

2. What is the standard deviation of $\bar{x}$?

3. Suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

4. Suppose instead we repeatedly took random samples of size 25. What proportion of the samples would we expect to have a sample mean $\bar{x}$ that lies between 106 and 110 g/L?

5. Again, suppose we take a large number of samples of size 75. What proportion of the samples would we expect to have a mean less than $\bar{x} = 103$?

6. If we repeatedly took samples of size 75, we would expect that, in 20% of the samples, $\bar{x}$ would be greater than _____?

7. After taking a sample of size 75, we found that the sample mean was $\bar{x} = 103$. Construct a two-sided 95% confidence interval for μ.

8. True or False. Based on the above interval, we can say that the probability that $\bar{x}$ lies in the interval is 0.95.

9. Suppose we were also interested in the mean of the highly right skewed indicator of iron absorption, ferritin. Compared to the relatively symmetrically distributed indicator hemoglobin, do you think a larger or smaller sample size would be required to apply the central limit theorem?

Hypothesis Testing with known Variance. Now, let's switch gears and assume that we didn't know the true population mean μ, and we only observed a sample of 25 school children in Delhi. Let's use the sample mean from this set of children to make inference about the true population mean μ. Assume that $\sigma = 12.5$ is known and that the sample size of 25 is large enough to use the Central Limit Theorem (i.e. base your inferences off of the normal distribution, not the t-distribution).

In this scenario, we can conduct a **one-sample Z-test** for inference about μ in a population with known variance σ^2. To test $H_0 : \mu = \mu_0$, we can use the test statistic:

$$Z^* = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

Under the null hypothesis, $Z^* \sim N(0, 1)$. So, we can use the standard normal distribution to calculate a p-value for this hypothesis test:

- For the one-sided test with alternative hypothesis $H_a : \mu > \mu_0$, we can calculate a p-value using the formula $p = P(Z > Z^*)$.
- For the one-sided test with alternative hypothesis $H_a : \mu < \mu_0$, we can calculate a p-value using the formula $p = P(Z \leq Z^*)$.
- For the two-sided test with alternative hypothesis $H_a : \mu \neq \mu_0$, we can calculate a p-value using the formula $p = 2 * P(Z < -|Z^*|)$.

Note: there is no command for directly conducting this one-sample Z-test in Stata. However, you can use the normal function in Stata to calculate the p-values.

1. In a sample of size 25, what is the value of the test statistic testing whether the mean hemoglobin level is equal to 108 g/L versus the alternative that it is not equal to 108 g/L, when $\bar{x} = 103$. Use a one-sample Z-test. What is the p-value?

2. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is less than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

3. In a sample of size 25, what is the value of the test statistic for testing whether the mean hemoglobin level in the population is equal to 108 g/L versus the alternative that it is greater than 108 g/L, when $\bar{x} = 103$. What is the p-value corresponding to this test?

Confidence intervals and testing with unknown variance. Suppose now that we are interested in the distribution of hemoglobin levels in Mumbai. We decide that it is unreasonable to extrapolate the Delhi results to Mumbai, and therefore the population standard deviation is unknown. We take a random sample of 15 children in Mumbai. The sample mean is $\bar{x} = 115$ g/L, with sample standard deviation $s = 10.2$ g/L. Assume hemoglobin levels in Mumbai are normally distributed (we could check this by looking at the distribution of hemoglobin levels in other similar populations).

1. Construct a two-sided 95% confidence interval for μ.

2. Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.05$ level?

   (a) yes (b) no (c) not enough information

   Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level?

   (a) yes (b) no (c) not enough information

   Using the above confidence interval, would we reject the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.1$ level?

   (a) yes (b) no (c) not enough information

3. Conduct a one-sample t-test in Stata to test the null hypothesis that the mean hemoglobin level is equal to 108 g/L, versus the alternative that the mean is not equal to 108 g/L, at the $\alpha = 0.01$ level.

   - What is your test statistic?

   - Under the null hypothesis, the test statistic follows a t-distribution with how many degrees of freedom?

   - What is your p-value?

- What do you conclude?

  (a) reject the null hypothesis, (b) accept the null hypothesis, (c) fail to reject the null hypothesis

Nursing Home Study. Suppose that 500 residents of a large nursing home are screened for hypertension. All residents with above a specified level are labeled as having hypertension. The following table displays the results of this study.

	Hypertension	No Hypertension	Total
Male	100	100	200
Female	100	200	300

1. Which of the following measures of association would be the most appropriate to use to describe the finding in this study?

    a. Cumulative Ratio
    b. Incidence Rate Ratio
    c. Prevalence Ratio

2. What is the prevalence of hypertension among all residents?

3. What is the prevalence odds ratio for having hypertension comparing male residents (exposed group) to female residents (non-exposed group)?

4. Which of the following is the **least** likely explanation for the value for the odds ratio reported in the previous answer?

    a. Men have a higher risk of developing hypertension than women.
    b. Women who develop hypertension live longer after being diagnosed than men.
    c. Men who develop hypertension are more likely to be residents of a nursing home than women who develop hypertension
    d. Hypertension is harder to detect in women than in men.

5. Is reverse causation a plausible explanation for the association seen in the study?

    a. Yes
    b. No

Bed Occupancy in Two Hospitals. An investigator performs a bed-occupancy survey on July 1, 2012 at two hospitals (Hospital A and Hospital B). Hospital A has 100 available beds. Hospital B is larger with 200 available beds. Hospital A reports that 80 of its beds are occupied on the day of the survey (occupancy prevalence = 0.80).

1. The investigator reports a Prevalence Odds Ratio of 2.25 when comparing the occupancy prevalence of Hospital B to that for Hospital A. What is the occupancy prevalence for Hospital B on that day?

Prevalence odds in Hospital A:
Prevalence odds ratio for Hospital B vs. Hospital A:
Prevalence odds in Hospital B:
Occupancy prevalence for Hospital B:

2. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. If the average length of stay is the same in each hospital then what is the incidence rate ratio for admission, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)?

a. 2.25
b. > 2.25
c. < 2.25
d. Cannot be determined from the information that is given

3. Suppose that the hospital administrators claim that occupancies for both hospitals are in steady state with the number of discharges on a given day equal to the number of discharges on that same day. Furthermore, suppose that the incidence rate ratio for admission is 1.5, comparing the admission rate (incidence rate) for Hospital B (exposed group) versus Hospital A (non-exposed group)? Would the average length of stay for patients in Hospital B be?

a. Less than that for Hospital A
b. Equal to that for Hospital A
c. Greater than that for Hospital A
d. Cannot be determined from the information that is given

4. If the investigator presented these results at a local conference, then what should he described as the design of this study?

a. Case Report
b. Ecologic Study
c. Cross Sectional Study