PHP 2500 Introduction to Biostatistics

Problem Set Four Solutions

---

1. For men in the group who will develop the disease, the cholesterol level, X, has a normal distribution with mean 244 and standard deviation 51. For men in the group who will not develop the disease, the cholesterol level, Y, has a normal distribution with mean 219 and standard deviation 41.

   A man will be predicted to get the disease if his level exceeds 260.

   (a) For a man who will develop the disease (i.e., one whose cholesterol level is represented by the random variable X) what is the probability that his level will exceed 260, so that he will be predicted (correctly) to develop the disease?

$$\begin{aligned} P(X > 260) &= P\left(\frac{X-244}{51} > \frac{260-244}{51}\right) \\ &= P(Z > 0.3137) \\ &= 0.3783 \end{aligned}$$

   (b) For one in the other group (who will not develop the disease, and whose level has distribution Y) what is the probability of predicting (wrongly) that he will develop the disease?

$$\begin{aligned} P(Y > 260) &= P\left(\frac{Y-219}{41} > \frac{260-219}{41}\right) \\ &= P(Z > 1) \\ &= 0.1587 \end{aligned}$$

   (c) For one in the group who will develop the disease (whose level has distribution X) what is the probability of failing to predict that he will develop the disease?

$$\begin{aligned} P(X < 260) &= P\left(\frac{X-244}{51} < \frac{260-244}{51}\right) \\ &= P(Z < 0.314) \\ &= 0.623. \end{aligned}$$

   (d) With a lower threshold, more men in both groups would be predicted to develop the disease (more would "test positive", and fewer would "test negative"), so the probability of false positive would increase, while the probability of false negative would decrease.

   (e) Not very useful-- a high proportion of the predictions will be wrong.

2. Let Y represent the albumin levels in cerebrospinal fluid among adults in the U.S. Now Y~N(29.5,9.25^2). Suppose we select repeated samples of size 20 from this population and calculate the mean for each sample.

(a) Here they are simply asking for the expected value of the sample mean: E[$\bar{Y}_{20}$]=E[Y]=29.5.

(b) The standard deviation of the sample mean is the standard error: sd[$\bar{Y}_{20}$]=sd[Y]/$\sqrt{20}$=9.25/4.47=2.068.

(c) The standard error (standard deviation of the sample means) is reduced by 1/$\sqrt{n}$ of the standard deviation.

(d) The distribution of sample means is **exactly** normal because the underlying distribution is normal (Why?). If the underlying distribution was non-normal, the Central Limit Theorem would tell us that the distribution of the sample means would be **approximately** normal in large samples.

(e) Standardizing the sample mean gives the result:

$$P(\bar{Y}_{20} > 33) = P(\sqrt{20}\,(\bar{Y}_{20} - 29.5)/9.25 > \sqrt{20}\,(33-29.5)/9.25)$$
$$= P(Z > 1.69)$$
$$= 0.0455$$

(f) Standardizing the sample mean gives the result:

$$P(\bar{Y}_{20} < 28) = P(\sqrt{20}\,(\bar{Y}_{20} - 29.5)/9.25 < \sqrt{20}\,(28-29.5)/9.25)$$
$$= P(Z < -0.7252)$$
$$= 0.2358$$

(g) Standardizing the sample mean gives the result:

$$P(29 < \bar{Y}_{20} < 31)$$
$$= P(\sqrt{20}\,(29-29.5)/9.25 < \sqrt{20}\,(\bar{Y}_{20} - 29.5)/9.25 < \sqrt{20}\,(31-29.5)/9.25)$$
$$= P(-0.2417 < Z < 0.7252)$$
$$= 1 - P(Z < -0.2417) - P(Z > 0.7252)$$
$$= 1 - 0.4052 - 0.2358 = 0.3590$$

3. $X \sim N(0,1^2)$.

(a) The distribution of sample means of size 10 that are drawn from X will (1) be **<u>exactly</u>** normally distributed, (2) have mean 0 $E[\overline{X}_{10}]=E[X]=0$, and (3) have variance $1/10=0.1$ (because $Var[\overline{X}_{10}]=Var[X]/10=0.1$).

(b) $P(\overline{X}_{10} > 0.6) = P(\sqrt{10}\,(\overline{X}_{10} - 0)/1 > \sqrt{10}\,(0.6-0)/1)$
$= P(Z > 1.8974)$
$= 0.0294$

(c) $P(\overline{X}_{10} < -0.75) = P(\sqrt{10}\,(\overline{X}_{10} - 0)/1 < \sqrt{10}\,(-0.75-0)/1)$
$= P(Z < -2.3717)$
$= 0.0089$

(d) Find ? when $P(\overline{X}_{10} > ?) = 0.2$
$P(\overline{X}_{10} > ?) = P(\sqrt{10}\,(\overline{X}_{10} - 0)/1 > \sqrt{10}\,(?-0)/1)$
$= P(Z > \sqrt{10}\,(?)) = 0.2$

But $P(Z > 0.84) = 0.2$, so $\sqrt{10}\,(?) = 0.84$ and $? = 0.2656$

(e) Find ? when $P(\overline{X}_{10} < ?) = 0.1$
$P(\overline{X}_{10} < ?) = P(\sqrt{10}\,(\overline{X}_{10} - 0)/1 < \sqrt{10}\,(?-0)/1)$
$= P(Z < \sqrt{10}\,(?)) = 0.1$

But $P(Z < -1.28) = 0.1$, so $\sqrt{10}\,(?) = -1.28$ and $? = -0.4048$

4. $X \sim N(3500,430^2)$ (approximately), where X is the birth weight for an infant whose gestational age is 40 weeks.

(a) $P(X < 2500) = P((X - 3500)/430 < (2500-3500)/430)$
$= P(Z < -2.33)$
$= 0.0099$

(b) Find ? when $P(X < ?) = 0.05$
$P(X < ?) = P((X - 3500)/430 < (?-3500)/430)$
$= P(Z < (?-3500)/430) = 0.05$

But $P(Z < -1.645) = 0.05$, so $(?-3500)/430 = -1.645$
and $? = 2792.65$

4. (c) The distribution of sample means of size 5 that are drawn from X will (1) be **<u>approximately</u>** normally distributed (because of the CLT), (2) have mean 3500 E[$\bar{X}_5$]=E[X]=3500, and (3) have variance $430^2/5=36980$ (because Var[$\bar{X}_5$]=Var[X]/5).

(d) Find ? when P($\bar{X}_5$ <?) = 0.1

P($\bar{X}_5$ < ?) = P($\sqrt{5}$ ($\bar{X}_5$ - 3500)/430 < $\sqrt{5}$ (?-3500)/430)

= P(Z < $\sqrt{5}$ (?-3500)/430) = 0.1

But P(Z < -1.645) =0.05, so $\sqrt{5}$ (?-3500)/430 = -1.645
and ? = 3183.66

(e) P($\bar{X}_5$ < 2500) = P($\sqrt{5}$ ($\bar{X}_5$ - 3500)/430 < $\sqrt{5}$ (2500-3500)/430)

= P(Z<-5.2)

= 0

⇨ (f) (Hmmm… This would make a good exam question!)
What is the probability that exactly one out of five babies have birth weight less than 2500 grams? Think Binomial! With the event of interest being a birth weight less than 2500 grams

We calculated in (a) that the probability of one baby having birth weight less that 2500 grams is P(X < 2500) =0.0099. So under a binomial set up we have n=5 and θ=0.0099 and we want to calculate $P(X=1)=\binom{5}{1}\theta^1(1-\theta)^4 = 5(0.0099)(0.9610)=0.0476$.

5. Let Y be the hemoglobin level for a female in the NHI survey. WE know that E[Y]=13.3 and $Var[Y]=1.12^2$.

(a) If repeated samples of 15 observations are collected, then by the CLT the distribution of the sample average $\bar{Y}_{15}$ will be approximately normal with mean E[$\bar{Y}_{15}$]=E[Y]=13.3 and Var[$\bar{Y}_{15}$]=Var[Y]/15=1.12^2/15=0.0836 (SD[$\bar{Y}_{15}$]=0.2891).
Thus

P(13< $\bar{Y}_{15}$ <13.6)=P((13-13.3)/0.2891<Z<(13.6-13.3)/0.2891)

=P(-1.0377<Z<1.0377) =1-2P(Z>1.04)=0.7016

5. (b) If repeated samples of 30 observations are collected, then by the CLT the distribution of the sample average $\bar{Y}_{30}$ will be approximately normal with mean $E[\bar{Y}_{30}]=E[Y]=13.3$ and $Var[\bar{Y}_{30}]=Var[Y]/30=1.12^2/30=0.0418$ ($SD[\bar{Y}_{30}]=0.2045$). Thus

$P(13<\bar{Y}_{30}<13.6)=P((13-13.3)/0.2045<Z<(13.6-13.3)/0.2045)$
$=P(-1.467<Z<1.467) =1-2P(Z>1.47)=0.8584$

(c) How large a sample size is required so that 95% of the sample means lie between +/- 0.20 grams of the true mean?

We want to calculate n so that $P(13.1<\bar{Y}_n<13.5)=0.95$, so the plan is to standardize and solve for n:

$0.95 = P(13.1<\bar{Y}_n<13.5)=$
$P(\sqrt{n}\,(13.1-13.3)/1.12 < \sqrt{n}\,(\bar{Y}_{20}-13.3)/1.12 < \sqrt{n}\,(13.5-13.3)/1.12)=$
$P(\sqrt{n}(-0.1786) <Z<\sqrt{n}(0.1786))$

But $P(-1.96<Z<1.96)=0.95$ so $\sqrt{n}(0.1786)=1.96$ and $n=120.43$. So 121 subjects are required.

(d) How large a sample size is required so that 95% of the sample means lie between +/- 0.10 grams of the true mean?

We want to calculate n so that $P(13.2<\bar{Y}_n<13.4)=0.95$, so the plan is to standardize and solve for n:

$0.95 = P(13.2<\bar{Y}_n<13.4)=$
$P(\sqrt{n}\,(13.2-13.3)/1.12 < \sqrt{n}\,(\bar{Y}_{20}-13.3)/1.12 < \sqrt{n}\,(13.4-13.3)/1.12)=$
$P(\sqrt{n}(-0.0893) <Z<\sqrt{n}(0.0893))$

But $P(-1.96<Z<1.96)=0.95$ so $\sqrt{n}(0.0893)=1.96$ and $n=481.73$. So 482 subjects are required.

6. Let X be the serum cholesterol level of an individual. Then $E[X]=211$ and $Var[X]=46^2$. We can to calculate the probability that the mean of sample of 25 observation is in the interval (195.9,226.1):

Thus we have $P(195.9 < \bar{X}_{25} < 226.1)=$
$P(\sqrt{25}\,(195.9-211)/46 < \sqrt{25}\,(\bar{X}_{25}-211)/46 < \sqrt{25}\,(226.1-211)/46)=$
$P(-1.64 <Z<1.64)=1-2P(Z>1.64) = 0.899$

The probability is 89.9%.

7. Sixty percent of the employees of a large heath care system were absent due to sickness for three or more days last year. Assume that the absentee rate due to illness will be the same this year and that we select a random sample of 150 employees.

(a) What is the probability that the sample proportion of employees absent three or more days will be between 0.5 and 0.65?

So we want to know P(0.5< $\hat{P}$ <0.65) where $\hat{P}$ is the sample proportion of employees absent three or more days. The underlying distribution is binomial, Y~Bin(n=150,θ=0.6) where Y represents the **number** of employees absent three or more days. Note that $\hat{P}$ =Y/n, so we know E[$\hat{P}$]=θ=0.6 and Var[$\hat{P}$]=Var[Y]/n=θ(1-θ)/n=0.6(0.4)/150=0.0016 (se=0.04). Because $\hat{P}$ is an average (of 1's and 0's) the central limit theorem applies and we can standardize:

P(0.5< $\hat{P}$ <0.65)=
=P((0.5-0.6)/0.04<($\hat{P}$ -θ)/sd($\hat{P}$)<(0.65-0.6)/0.04)
=P(-02.5<Z<1.25) =1-0.1056-0.0062=0.8882

(b) What is the probability that the sample proportion will exceed 0.7?
Same reasoning as above applies:
P($\hat{P}$ >0.7)=
=P(($\hat{P}$ -θ)/sd($\hat{P}$)<(0.7-0.6)/0.04)
=P(Z>2.5) = 0.0062

(c) What is the expected number in the sample who will be absent three or more days due to illness?
E[Y]=nθ=90, so 90 people are expected to be absent for at least 3 days.

8. The average rate of gun deaths in providence is 8.7 per month. Suppose we track the number of gun deaths by month over a two-year period.

(a) What is the probability that the average gun death rate over the two-year period will be between 7.5 and 10 per month?

Let X be the number of gun deaths in a one month period. Then X~Poiss(λ=8.7) and we are observing $X_1,X_2,\ldots,X_{24}$. Hence E[$\bar{X}_{24}$]=E[X]= λ=8.7 and Var[$\bar{X}_{24}$]=Var[X]/n=λ/n=8.7/24=0.363 (se=0.6021). Because $\bar{X}_{24}$ is an average the central limit theorem applies and we can standardize:

P(7.5< $\bar{X}_{24}$ <10)=
= P((7.5-8.7)/0.6021<($\bar{X}_{24}$ -λ)/sd($\bar{X}_{24}$)<(10-8.7)/0.6021)
= P(-1.99 < Z < 2.16) =1-0.0228-0.0154 = 0.9618

(b) What is the probability that the sample average will exceed 12 deaths per month in just one year?

Same reasoning as above except only 12 observations are being collected: Hence E[$\bar{X}_{12}$]=E[X]= λ=8.7 and Var[$\bar{X}_{12}$]=Var[X]/n=λ/n=8.7/12=0.725 (se=0.8515). Because $\bar{X}_{12}$ is an average the central limit theorem applies and we can standardize:

P($\bar{X}_{12}$ > 12)=
= P(($\bar{X}_{12}$ -λ)/sd($\bar{X}_{12}$)>(12-8.7)/0.8515)
= P(Z > 3.8755) = 0

(c) What is the total expected number of gun deaths over the two year period?

The expected number of deaths in one month is E[X]= λ=8.7, so the expected number of deaths over a two year period is 8.7(24) = 208.8.

9. The Study habits portion of the survey of Study Habits and Attitudes (SSHA) psychological test consists of two sets of questions. One set of questions measures "delay avoidance" (procrastination) and the other measures "work methods". A subject's study habits score is the sum X+Y of the delay avoidance score X and the work methods score Y. The distribution of X in a broad population of students at Schools of Public Health is N(25,10) and the distribution of Y in the same population is N(30,9).

   (a) If a subject's X and Y scores are independent, what is the distribution of the study habits score X+Y?

   Because X and Y are normal X+Y~N(55,19).
   (Var[X+Y]=Var[X]+Var[Y]=10+19 because of independence.)

   (b) Assuming the scores are independent, what percentage of the population has a study habits score of 60 or higher?

   $P(X+Y>60) = P(Z>(60-55)/4.36)=P(Z>1.15)=0.1251$

   (c) In fact, the X and Y score are strongly correlated with a correlation of 0.7. In this case what is the effect on the expected value and variance you found in (a)?

   $E[X+Y] = E[X]+E[Y]=25+30=55$
   $Var[X+Y] = Var[X] + Var[Y] + 2cov(X,Y)$
   $= 10 + 9 + 2(0.7)(\sqrt{10}\sqrt{19})$
   $= 32.28$

   (d) Assuming the study habits score has a normal distribution with the mean and variance found in (c), what is the proportion in the population with a study habits score exceeding 60?

   $P(X+Y>60) = P(Z>(60-55)/5.68)=P(Z>0.88)=0.1894$

   (e) Repeat (d) with a correlation of –0.7.

   $E[X+Y] = E[X]+E[Y]=25+30=55$
   $Var[X+Y] = Var[X] + Var[Y] + 2cov(X,Y)$
   $= 10 + 9 - 2(0.7)(\sqrt{10}\sqrt{19})$
   $= 5.718$
   Thus,
   $P(X+Y>60) = P(Z>(60-55)/2.39)=P(Z>2.09)=0.0183$

10. Two investigators independently estimate the mean of a population. There estimates are $\overline{X}$ (std.err. =0.3) and $\overline{Y}$ (se=1.2) respectively. As the project statistician you are asked to combine the two estimates to obtain an overall estimate (this is one of the problems addressed by Meta-Analysis which is more appropriately called combining information). Three alternatives proposed by colleagues are:

a. Simply use $\overline{X}$ because it has the smallest std.err.

b. Average the two estimates $(\overline{X} + \overline{Y})/2$

c. Use $0.94\overline{X} + 0.06\overline{Y}$ because it came to you in a dream:

$$\frac{(1.2)^2}{(0.3)^2 + (1.2)^2}\overline{X} + \frac{(0.3)^2}{(0.3)^2 + (1.2)^2}\overline{Y} = 0.94\overline{X} + 0.06\overline{Y}$$

(We will discuss this estimate in more detail later.)

(a) Show that each of the three proposed estimates has expected value equal to μ, the mean of the population.

$E[\overline{X}]=E[X]=\mu$
$E[(\overline{X} + \overline{Y})/2]=(E[X]+E[Y])/2=\mu$
$E[0.94\overline{X} + 0.06\overline{Y}]=0.94\ E[X]+0.06\ E[Y]=0.94\mu+0.06\mu=\mu$

(b) Find the variance of each proposed estimate. Which is preferred in the sense of having the smallest variance?

$Var[\overline{X}]=0.3^2=0.09$ and $Var[\overline{Y}]=1.2^2=1.44$
$Var[(\overline{X} + \overline{Y})/2]=(Var[\overline{X}]+Var[\overline{Y}])/4=(0.09+1.44)/4=0.3825$
$Var[0.94\overline{X} + 0.06\overline{Y}] = 0.94^2\ Var[\overline{X}]+0.06^2\ Var[\overline{Y}]$
$= 0.8836(0.09) + 0.0036(1.44)=0.0847$

The point of this exercise is to show that weighted averages can be better in the sense that their variance is smaller than the equally weighted average or the simple average. In fact, by weighting by fraction of total variance, as was done for estimator (c), you will always get the estimator with the smallest variance.

11. Anatomy of a box plot: suppose that X is normally distributed, having mean μ and variance σ^2.

(a) Show that the inter-quartile range of X is 1.34σ.
The inter-quartile range is the middle 50% of the distribution, defined in class as 75 centile – 25 centile.
That is IQR=B-A where B and A are defined by:

$$P(X \le A) = 0.25 \quad \text{and} \quad P(X \le B) = 0.75$$

To find A and B we standardize:
$P(X<A)=P(Z<(A-\mu)/\sigma)=0.25$ so $(A-\mu)/\sigma = -0.667$
and $A = \mu - 0.667\sigma$

$P(X>B)=P(Z>(B-\mu)/\sigma)=0.25$ so $(B-\mu)/\sigma = 0.667$
and $B = \mu + 0.667\sigma$

$IQR=B-A=\mu + 0.667\sigma - (\mu - 0.667\sigma) = 1.34\sigma$

(b) Using the result in (a) find the probability of exceeding the "upper fence" defined by

$$upf = B + \frac{3}{2}(1.34\,\sigma)$$

$P(X > upf) = P(Z > (B+1.5(1.34\sigma)-\mu)/\sigma)$
$= P(Z > (\mu + 0.667\sigma+1.5(1.34\sigma)-\mu)/\sigma)$
$= P(Z > 2.67) = 0.0038$

(c) Using the result in (a) find the probability of exceeding the "lower fence" defined by

$$lof = A - \frac{3}{2}(1.34\,\sigma)$$

$P(X < lof) = P(Z < (A-1.5(1.34\sigma)-\mu)/\sigma)$
$= P(Z < (\mu - 0.667\sigma-1.5(1.34\sigma)-\mu)/\sigma)$
$= P(Z < -2.67) = 0.0038$

(d) Using the result of (b) and (c) explain why observation beyond the "fences" may be considered as exceptional or unusual values.

The probability of observing an "exceptional value" is very small indicating that something else might be going on.