STAT 13, section 1, Winter 2012, UCLA Statistics

HW 5; Problem Solution

- (**HW.5.1**) Suppose we are to draw a random sample of *five individuals* from a large population in which 40% of the individuals are mutants. Let *p^* represent the *proportion of mutants in the sample* (sample proportion, an estimate of the population proportion of mutants).
  - Use the Binomial Distribution to determine the probability that *p^* will be equal to:

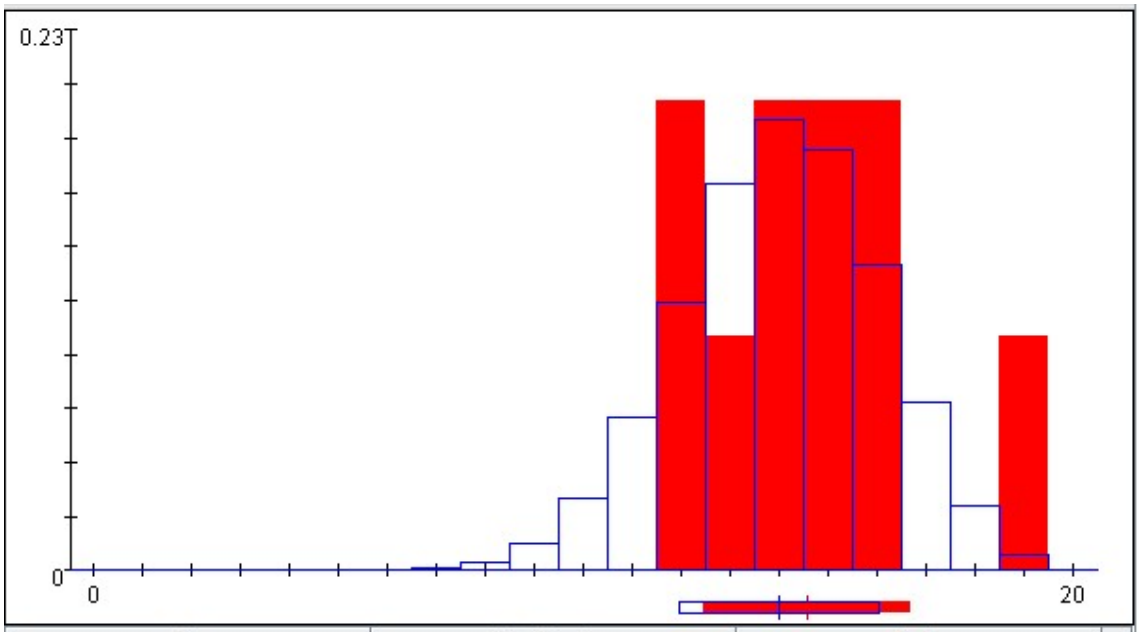
$$T \sim Bin(5,0.4), \hat{P} = \frac{T}{5} \sim \frac{1}{5} Bin(5,0.4)$$

T	0	1	2	3	4	5
$\hat{P}$	0	0.2	0.4	0.6	0.8	1
Prob	$C_5^0 0.4^0 0.6^5$	$C_5^1 0.4^1 0.6^4$	$C_5^2 0.4^2 0.6^3$	$C_5^3 0.4^3 0.6^2$	$C_5^4 0.4^4 0.6^1$	$C_5^5 0.4^5 0.6^0$
	0.07776	0.25920	0.34560	0.23040	0.07680	0.01024

Prob of $\hat{P}$ equals:

- 0 0.078
- 0.1 0
- 0.5 0
- 0.6 0.23
- 0.7 0
- 1.0 0.01

  - Use Binomial Coin Experiment to run 10 experiments each consisting of tossing 20 coins (with p=0.7).



X	Distribution	Data
0	0	0
1	0	0
2	0	0
3	0	0
4	0.00001	0
5	0.00004	0
6	0.00022	0
7	0.00102	0
8	0.00386	0
9	0.01201	0
10	0.03082	0
11	0.06537	0
12	0.1144	0.2
13	0.16426	0.1

- Compare the empirical probabilities (the third column in the results table) to the exact/theoretical probabilities (second column in the results table).
- How close are the values in columns 2 and 3?

  Not very close.

- Are they supposed to be similar? Would they become more or less similar if we increase the number of trials to 100, for each experiment?

  Yes, but the number of experiments here is small, so we can't really see the exact similarity between these two.

  No. We need to increase the number of experiments to make these two columns closer.

- Would the values in columns 2 and 3 become more or less similar if we run 100 cumulative experiments (each of 20 coins)?

Yes.

- (**HW.5.2**) An important indicator of lung function is forced expiratory volume (FEV), which is the volume of air that a person can expire in one second. Dr. Jones plans to measure FEV in a random sample of *n* young women from a certain population and to use the sample-mean as an estimate of the population-mean. Let *E={event that Dr. Jones sample-mean is within +/-100 mLi of the true (unknown) population-mean}*. Assume that the population is *Normal(mean=3,000 mLi, variance=400^2 mLi)*. Find *P(E)*, if:

$$\bar{X} \sim N(3000, \frac{400^2}{n})$$

Calculate P(E) = $P(|\bar{X} - 3000| \leq 100) = P(\frac{|\bar{X}-3000|}{\sqrt{\frac{400^2}{n}}} = |Z| \leq \frac{100}{\sqrt{\frac{400^2}{n}}})$

- *n=20*

$$P(E) = P\left(\frac{|\bar{X} - 3000|}{\sqrt{\frac{400^2}{n}}} = |Z| \leq \frac{100}{400\sqrt{\frac{1}{20}}} = 1.12\right) = 0.74$$

- *n=80* (quadruple expansion of the sample size)

$$P(E) = P\left(|Z| \leq \frac{100}{400\sqrt{\frac{1}{80}}} = 2.236\right) = 0.975$$

- How does *P(E)* depend on the sample size? Does *P(E)* increase, decrease or stay unchanged with increase of *n*?

P(E) increases with the sample size.

- (**HW.5.3**) A certain cross between sweet-pea plants will produce progeny that are either purple flowered of white flowered; the probability of purple-flowered plant is p=9/16. Suppose n progeny are to be examined, and let *p^* be the sample proportion of purple-flowered plants. It may happen by chance that *p^* would be closer to 1/2 than to 9/16.

Use Normal-approximation without continuity correction to find the probability that this misleading event will occur if

$$\hat{p} \sim N(p, \frac{p(1-p)}{n}), \text{ if n is large enough}$$

(1/2 + 9/16)/2 = 17/32

{Event: $\hat{p}$ closer to 1/2 = 8/16} = {Event: $\hat{p}$ <17/32=0.53125}

{Event: $\hat{p}$ closer to 9/16} = {Event: $\hat{p}$ >17/32=0.53125}

- n=1

$$\hat{p} = X_1 \sim Ber(p) = Ber(\frac{9}{16})$$

P($\hat{p}$ closer to 1/2)=P($\hat{p}$ <17/32)=P(X_1<17/32) = P(X_1=0) = 7/16

P($\hat{p}$ closer to 9/16)=P($\hat{p}$ >17/32)=P(X_1>17/32) = P(X_1=1) = 9/16

$\hat{p}$ will have slightly more chance to be closer to 9/16

$$N\left(p, \frac{p(1-p)}{n}\right) = N(0.5625, 0.246)$$

- n=64

$$\hat{p} \sim Bin\left(64, \frac{9}{16}\right) \to N\left(\frac{9}{16}, \frac{9}{16} * \frac{7}{16} * \frac{1}{64}\right) = N(0.5625, 0.00385)$$

P($\hat{p}$ closer to 1/2)=P($\hat{p}$ <17/32=0.53125)=0.31

P($\hat{p}$ closer to 9/16)=P($\hat{p}$ >17/32=0.53125)=0.69

It will have once more chance to be closer to 9/16.

- n=320

$$\hat{p} \sim Bin\left(64, \frac{9}{16}\right) \to N\left(\frac{9}{16}, \frac{9}{16} * \frac{7}{16} * \frac{1}{320}\right) = N(0.5625, 0.00077)$$

P($\hat{p}$ closer to 1/2)=P($\hat{p}$ <17/32=0.53125)=0.13

P($\hat{p}$ closer to 9/16)=P($\hat{p}$ >17/32=0.53125)=0.87

It has much more chance to be closer to 9/16.

- (**HW.5.4**) In a certain lab population of mice, the weights at 20 days of age follow approximately Normal distribution with mean weight=8.3g and standard deviation=1.7g. Suppose many litters of 10 mice each are to be weighted. If each litter can be regarded as a random sample from the population, what percentage of litters will have total weight of 90g or more? (Hint: How is the total weight of a litter related to the mean weight of its members?)

$$X \sim N(8.3, 1.7^2),$$

$$Y = X_1 + \cdots + X_{10} \sim N(83, 10 * 1.7^2) = N(83, 28.9)$$

Let E be a random variable note the total weight less than or no less than 90g:

$$P(E = 1) = P(\text{This litter has total weight no less than 90g}) = P(Y \geq 90) = 0.1$$

$$P(E = 0) = P(\text{This litter has total weight less than 90g}) = P(Y < 90) = 0.9$$

So E~Bernoulli(p=0.1).

Right now we are sampling $Y_1, Y_2, \ldots, Y_n$, corresponding to $E_1, E_2, \ldots E_n$. Here n is large, so we can expect the percentage to have a normal distribution based on CLT:

$$\hat{p} = \frac{1}{n}(E_1 + E_2 + \cdots + E_n) \sim N\left(0.1, \frac{0.1 * 0.9}{n}\right) = N(0.1, \frac{0.09}{n})$$

The expected percentage will be 0.1, while the real observed percentage will follow the distribution above.

- (**HW.5.5**) Six healthy three-year-old female Suffolk sheep were injected with the antibiotic Gentamicin at a dosage of 10 mg/kg body weight. Their blood serum concentrations (μg/mLi) of Gentamicin 1.5 hours after injection were as follows: {29 26 30 31 23 28}. For these data, the mean is 27.8 and the standard deviation is 2.9.

We use a normal distribution for the sample mean.

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{6}\right) \approx N\left(\mu, \frac{\hat{\sigma}^2}{6}\right).$$

- Construct a 95% confidence interval for the population mean

$$\hat{\mu} \pm z_{\frac{0.05}{2}} * \frac{2.9}{\sqrt{6}} = 27.8 \pm 1.96 * \frac{2.9}{\sqrt{6}}$$

[27.8-2.32, 27.8 + 2.32] = [25.48, 30.12]

Using t-test:

$$\hat{\mu} \pm t_{\frac{0.05}{2},5} * \frac{2.9}{\sqrt{6}} = 27.8 \pm 2.57 * \frac{2.9}{\sqrt{6}}$$

[27.8-3.04, 27.8 + 3.04] = [24.76, 30.84]

- Define in words the population mean that you estimated above.

If we repeated sampling 6 data points here, there will be 95% chance that our confidence intervals will cover the true mean, and 5% chance missing the true mean.

This interval [25.48, 30.12] is one realization of possible confidence intervals mentioned above. With 95% chance it is one of the covering intervals, but also with 5% it is among those missing intervals.

- Is it typical for the 95% confidence interval constructed in the first part to nearly contain all of the observations?

It is not typical. If we sample more data points, say 6,000 instead of only 6, then the confidence interval will be very narrow, and lots of observations will be locating out of the confidence intervals.

- (**HW.5.6**) Human beta-endorphin (HBE) is a hormone secreted by the pituitary gland under conditions of stress. A researcher conducted a study to investigate whether a program of regular exercise might affect the resting (unstressed) concentration of HBE in the blood. He measured blood HBE levels, in January and again in May, in ten participants in a physical fitness program. The results were as shown in the table.

HBE Level (pg/mLi)			
Participant	**January**	**May**	**Difference**
1	*42*	*22*	*20*
2	*47*	*29*	*18*
3	*37*	*9*	*28*
4	*9*	*9*	*0*
5	*33*	*26*	*7*
6	*70*	*36*	*34*
7	*54*	*38*	*16*
8	*27*	*32*	*-5*
9	*41*	*33*	*8*
10	*18*	*14*	*4*
Mean	**37.8**	**24.8**	**13.0**
SD	**17.6**	**10.9**	**12.4**

  - Conduct a 95% confidence interval for the population mean difference in HBE levels between January (unstressed level) and May (HBE level possibly perturbed by exercise). (Hint: You need to use only the values in the right-hand column.)

    By normal test:

$$\hat{\mu} \pm z_{0.05/2} * \frac{12.4}{\sqrt{10}} = 13.0 \pm 1.96 * \frac{12.4}{\sqrt{10}} = 13.0 \pm 7.69$$

$$[13.0 - 7.69, 13.0 + 7.69] = [5.31, 20.69]$$

    By t-test:

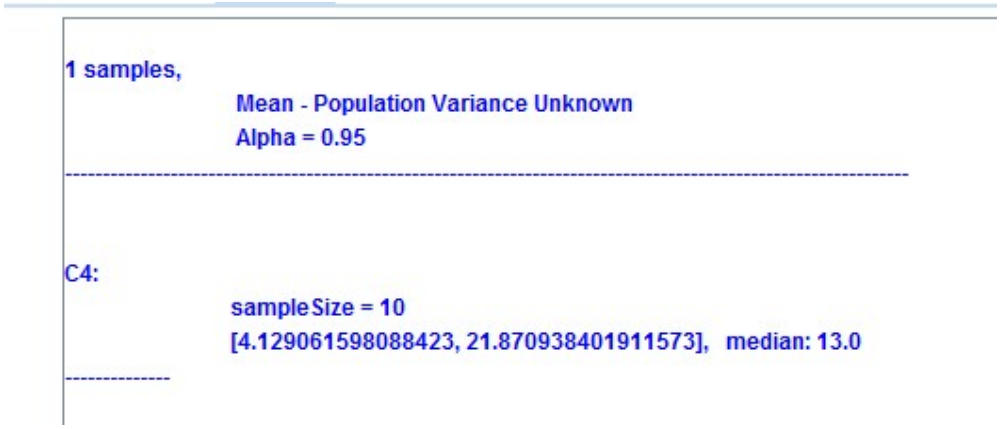
$$\hat{\mu} \pm t_{\frac{0.05}{2},9} * \frac{12.4}{\sqrt{10}} = 13.0 \pm 2.26 * \frac{12.4}{\sqrt{10}} = 13.0 \pm 8.86$$

$$[13.0 - 8.86, 13.0 + 8.86] = [4.14, 21.86]$$

- Interpret the confidence interval from the first part - that is to explain what the confidence interval tells you about the HBE levels.

If I repeated these experiments by asking N groups of ten participants to do the same test, I can construct N confidence intervals. There are about 95% of these N confidence intervals that will cover the true population mean difference. $[13.0 - 8.86, 13.0 + 8.86] = [4.14, 21.86]$ is just one realization of those confidence intervals, it might cover the true mean, or not.

- Use the SOCR Confidence Interval Analysis applet to compute the interval for the population mean difference between January and May (paste in all 4 columns in the data tab, choose mean unknown variance in the CI Settings tab, map the 4th column, difference, as the data column in the Mapping tab, and finally click Calculate button). Compare these results to your manual calculations.



Our results:

Z test:

$$[13.0 - 7.69, 13.0 + 7.69] = [5.31, 20.69].$$

t-test:

$$[13.0 - 8.86, 13.0 + 8.86] = [4.14, 21.86].$$