

**Part I**

In the entire population of water supplies in WA, the concentration of fluoride is known to be approximately normally distributed, with mean 3 ppm and standard dev 1 ppm.

a. This question reviews our use of the cumulative normal probability distribution where the random variable  $X$  is the concentration of fluoride in the water supply. We want to calculate the probability of a concentration greater than 2 ppm:

$$P(X > 2) = P(Z > (2-3)/1) = P(Z > -1) = .5 + P(-1 < Z < 0) = .5 + .3413 = 0.841$$

Thus, we can say that about 84% of the water supplies will have a concentration of greater than 2 ppm of fluoride.

Next, what is the chance that a random water supply contains 1 to 2 ppm? We want:

$P(1 < X < 2) = P(X < 2) - P(X < 1)$ . Well,  $P(X < 2) = 0.1587$  from above.  $P(X < 1) = P(Z < (1-3)/1) = P(Z < -2) = 0.0228$ . And  $0.1587 - 0.0228 = 0.1359$ . Thus, about 13.6% of the water supplies will have a concentration between 1 and 2 ppm of fluoride.

b. Now we need:  $P(X_1 > 2 \text{ and } X_2 > 2)$  (remember that the probability of independent events both occurring is the product of their individual probabilities)  $= P(X_1 > 2)P(X_2 > 2) = (0.841)(0.841) = 0.707$ , thus about 70.7% of the time two randomly selected water supplies will both contain more than 2 ppm.

Next we calculate  $P(1 < X_1 < 2 \text{ and } 1 < X_2 < 2) = P(1 < X_1 < 2)P(1 < X_2 < 2) = (0.1359)(0.1359) = 0.0185$ .

In both cases in part b. the probabilities are a lot lower than the probability that one random water supply will satisfy the condition given, since it is less likely to happen that two random draws from the population will both satisfy the condition.

c. Now we average the two concentrations and create a new random variable:

$\bar{X} = (X_1 + X_2)/2$ .  $\bar{X}$  has a new mean and standard deviation which can be calculated according to the rules of expectation from about a linear combination of independent random variables.

$$E[X+Y]=E[X]+E[Y], \text{ etc.}$$

$$\text{The mean is } E[\bar{X}] = E[(X_1 + X_2)/2] = (1/2)(E[X_1] + E[X_2]) = (1/2)(3 + 3) = 3.$$

$$\text{The var is } V[\bar{X}] = V[(X_1 + X_2)/2] = (1/4)(V[X_1]+V[X_2]) = (1/4)(2V[X]) = 0.5, \text{ or it equals } V(X)/n = 1/2 = .5.$$

So, just as we have seen in class the sample mean is on average equal to the population mean, but has less variance than random selections from the population. The standard deviation is the square root of the variance, so it equals .707.

$$\text{Finally, } P(\bar{X} > 2) = 1 - P(\bar{X} < 2) = 1 - P(Z < (2-3)/0.707) = 1 - P(Z < -1.414) \\ = 1 - 0.0793 = 0.9207$$

$$\text{And } P(1 < \bar{X} < 2) = P(\bar{X} < 2) - P(\bar{X} < 1) = 0.0793 - 0.0023 = 0.077$$

The probability that the average of two random water supply concentrations is over 2 ppm is 92.07%, higher than for one water supply since we expect the average to fall close to the mean concentration of 3

ppm. In other words, the chance that the average of the two would fall below 2 ppm (away from the mean) is less than it would be for a single draw. The probability that the average will be between 1 and 2 ppm is only about 7.7% since once again the average of the two supplies would be expected to fall close to the mean of 3 ppm.

d. Calculate mean and standard deviation for a random sample of 10 water supplies.

$E[\bar{X}] = 3$  and  $V[\bar{X}] = s^2/n = 1/10 = 0.1$ . Thus  $SD[\bar{X}] = 0.32$ . Wow, the standard deviation and variance for the mean of 10 samples are small!!

$P(\bar{X} > 2) = 1 - P(Z < (2-3)/0.32) = 1 - P(Z < -3.13)$ .  $P(Z < -3.13)$  is basically off of the charts, i.e., it is very close to 0, thus  $1 - P(Z < -3.13)$  is very close to 1. This makes sense because the average concentration in 10 samples should be *extremely close* to the population mean which is 3 ppm (included in the range that is above 2 ppm).

And furthermore, there is an extremely tiny probability of an average concentration for 10 samples of between 1 and 2 ppm for the same reason.

e. If you know that a sampling distribution where  $n=10$  has a mean of 3ppm and a standard deviation (standard error) of 1ppm, there are a few things you can say about the population mean:

$\mu = 3$  – since the mean of a sampling distribution is equal to the population mean.

$$s = 1\text{ppm} * \sqrt{10} = 3.16$$

If the sample size is 40, but the sampling distribution still has a standard error of 1ppm, the standard deviation of the population must be larger. We can solve for it:

$$\text{Std.Error} = \frac{s}{\sqrt{n}} \dots \text{so } s = \text{Std.Error} \sqrt{n}$$

$$s = 1 \times \sqrt{40} = 6.32$$

Therefore, if the sample size gets bigger while the standard error of the sampling distribution stays the same size, it must mean that the population standard deviation grew. Increases in sample sizes reduce the standard error of sampling distributions, but the population standard deviation also has an important effect on the width of sampling distributions.

## Moving from Populations to Samples

### Part II

1a. We want the 95 percent confidence interval. We are given the mean and the sample standard deviation.  $n=21$ . So, we need to use the t-distribution because the population SD is unknown and the sample size is small. If the CI is 95% then  $\alpha=5\%=.05$  and  $\alpha/2=.025$ .  $t_{\alpha/2}=2.086$  with 20 degrees of freedom ( $n-1=21-1=20$  df). We want this confidence interval:

$$P(\bar{x} - 2.086SE < \sigma < \bar{x} + 2.086SE) = .95$$

Next, we'll calculate the estimate of the standard error.  $SE = \frac{s}{\sqrt{n}} = \frac{3.22}{\sqrt{21}} = .70$

Then, we'll plug in SE estimate and sample mean:

$$P(\bar{x} - 2.086SE < \mu < \bar{x} + 2.086SE) = .95$$

$$= P[52.6 - (2.086)(.7) < \mu < 52.6 + (2.086)(.7)] = P[51.14 < \mu < 54.06] = .95$$

so,  $\bar{x} \pm 2.086SE = [51.14, 54.06]$

1b. The mean pulse rate for all U.S. adult males is 72 heartbeats per minute. 95 percent of the time, the heart rate for an U.S. adult man will fall in the range from 51.14 to 54.06 beats per minute. Only 5% percent of the time will it fall out of the range. So, we can say that it appears that jogging 15 miles per is associated with a reduced heart rate in adult males.

1c. We need to assume that the heart rate is normally distributed in the population.

2. We want to know if these sample sizes are large enough to construct a confidence interval for the true proportion of employees at each fitness level exhibiting signs of stress. First, we need to think about how close we need to be to "true". Do we need to be within  $\pm .01$ ?  $\pm .02$ ?  $\pm .03$ ? What would meet our needs? Second, will a 95% confidence interval provide us with enough confidence or should the interval be larger, say 99%? Likely being 99% confident is fine. In order to figure out the sample size, we need an estimate of the variance. The table below presents that estimate,  $\hat{p}\hat{q}$

Fitness Level	Sample Size	Proportion with Signs of Stress $\hat{p}$	$\hat{q}$	$\hat{p}\hat{q}$ (Var)	$\sqrt{\hat{p}\hat{q}}$ (est. SD)
Poor	242	.155	.845	.131	.362
Average	212	.133	.867	.115	.339
Good	95	.108	.892	.096	.310

2a. If we choose a 95% confidence interval,  $z_{\alpha/2}=1.96$ .

To see how close we could get to the true estimate, we can use a transformation of the sample size

estimation equation:  $B = z_{\alpha/2} \frac{S}{\sqrt{n}}$

$$B_{poor} = 1.96 \frac{.362}{\sqrt{242}} = .046 \quad B_{average} = 1.96 \frac{.339}{\sqrt{212}} = .046 \quad B_{good} = 1.96 \frac{.310}{\sqrt{95}} = .062$$

So, yes, the samples are large enough to construct a confidence interval for the true proportion. For poor and average, we can get within  $\pm .05$  with 95% confidence, and for good fitness we can get within  $\pm .06$ . If we want make an estimate that will be closer to the actual value (narrow the confidence interval) we need to increase the sample size or increase the confidence interval.

$$2b. \quad P(\hat{p}_{poor} - 1.96SE < p_{poor} < \hat{p}_{poor} + 1.96SE) = .95$$

$$= P[.155 - (1.96)\left(\sqrt{\frac{.131}{242}}\right) < p_{poor} < .155 + (1.96)\left(\sqrt{\frac{.131}{242}}\right)] = P[.109 < p_{poor} < .201] = .95$$

$$\text{so, } \hat{p}_{poor} \pm 1.96SE = [.109, .201]$$

$$P(\hat{p}_{average} - 1.96SE < p_{average} < \hat{p}_{average} + 1.96SE) = .95$$

$$= P[.133 - (1.96)\left(\sqrt{\frac{.115}{212}}\right) < p_{average} < .133 + (1.96)\left(\sqrt{\frac{.115}{212}}\right)] = P[.087 < p_{average} < .179] = .95$$

$$\text{so, } \hat{p}_{average} \pm 1.96SE = [.087, .179]$$

$$P(\hat{p}_{good} - 1.96SE < p_{good} < \hat{p}_{good} + 1.96SE) = .95$$

$$= P[.108 - (1.96)\left(\sqrt{\frac{.096}{95}}\right) < p_{good} < .108 + (1.96)\left(\sqrt{\frac{.096}{95}}\right)] = P[.046 < p_{good} < .170] = .95$$

$$\text{so, } \hat{p}_{good} \pm 1.96SE = [.046, .170]$$

2c. We are 95 percent confident (given our sample sizes for each) group that, of those with poor fitness levels, between 11% and 20% will show signs of stress. Of those with average fitness levels between 9% and 18% will show signs of stress. Of those with good fitness levels between 5 and 17 % will exhibit signs of stress. These confidence intervals overlap, leading us to question whether there are real differences in the rates for the groups. However, all these sample sizes differ, the distance from the mean differs, and these results do depend on the sample size. If we increased the sample size for those in good health, the 95% confidence interval would decrease in size. Therefore, although they are all 95% confidence intervals, it is difficult to compare them.

2d. We'd like to get the sample size to be 95% and within .01. So, we use our equation:

$$n = \frac{z_{\alpha/2}^2 pq}{D^2}, \text{ where } z_{\alpha/2} = 1.96 \text{ and } D = .01 \text{ } pq \text{ is calculated in the table above.}$$

So  $n = \frac{1.96^2 (.096)}{.01^2} = 3688$  This is a larger sample size due to the smaller distance we can tolerate from the mean.

### Part III

There are no unique answers for these problems so this answer key will walk through a generic set for those presented in the assignment sheet.

**1.** The hypothesis was that public transportation users would have lower than average personal wage earnings because higher wage earners would use their income to "buy" convenience (parking, cars, etc.)

**2.** Below is the printout of COMPARE MEANS.

2001 Personal Wage Earnings

Public Transit	Mean	N	Std. Deviation	Std. Error of Mean
Non-Public Transit Method	39961.26	2937	47161.299	870.229
Public Transit Method	34223.07	138	24975.522	2126.058
Total	39703.74	3075	46406.361	836.864

To construct a confidence interval about the proportion of cases in the public transportation category we need to calculate an estimate: The point estimate is just  $138/3075 = 0.05$  (5% of people use public transportation).

To get the standard deviation,  $S$ , estimate for a proportion, we will first find the variance, then take the square root.  $\text{Var}(\text{proportion}) = p(1-p)/n = .05*(1-.05)/3075 = (.0475)/3075 = .0001545$  and its square root equals 0.004.  $N$  here is 3075 because the mean is a proportion of the 3075 cases. Since this is a large sample, we just use the normal tables. For a 99% confidence interval (0.5% in each tail) we find that  $Z = 2.58$  (or 2.576 to be exact).

Now, we have all of the tools, the sample proportion ( $\hat{p}$ ), and its standard deviation ( $s$ )

$$P((P-2.58*S < p < P+2.58*S) = 0.99$$

$$P((0.05-2.58(0.004)) < p < (0.05+2.58(0.004))) = 0.99$$

$$P((0.05 - 0.01032) < p < (0.05 + 0.01032)) = 0.99$$

$$P((0.03968) < p < (0.06032)) = 0.99$$

So I can say with a 99% confidence level that between 3.9% and 6.0% of households in WA use public transportation to commute to work.

**3.** To find confidence intervals for the mean outcome for each category:

get SE for each category (REMEMBER THAT SPSS GIVES YOU THE SE BUT ON A QUIZ YOU MIGHT BE ASKED TO CALCULATE IT YOURSELF!)

$$\text{For non-public transportation} = 0 \quad SE = S/\sqrt{n} = 47161/\sqrt{2937} = 870.229$$

$$\text{For public transportation} = 1 \quad SE = S/\sqrt{n} = 24975/\sqrt{138} = 2126.06$$

get t-statistic 95% and 90% for each category

$$\text{For non-public transportation} = 0, \text{ DF} = 2937-1=2936$$

$$t_{0.025} = 1.96 \quad t_{0.05} = 1.64$$

$$\text{For public transportation} = 1, \text{ DF} = 138-1 = 137$$

$$t_{0.025} = 1.96 \quad t_{0.05} = 1.64$$

Note that as  $n$  is large in both cases we get the same results by using the normal table.

construct CI for transportation = 0

$$P(\bar{X} - 1.64S/\sqrt{n} < \mu \leq \bar{X} + 1.64S/\sqrt{n}) = 0.9$$

$$=P[39961-1.64(870.229) < \mu < 39961+1.64(870.229)] = 0.9$$

$$=P[38533 < \mu < 41388] = 0.9$$

$$P(\bar{X} - 1.96S/\sqrt{n} < \mu < \bar{X} + 1.96S/\sqrt{n}) = 0.95$$

$$= P[39961-1.96(870.229) < \mu < 39961+1.96(870.229)] = 0.95$$

$$=P[38225 < \mu < 41666] = 0.95$$

construct CI for transportation =1:

$$P(\bar{X} - 1.64S/\sqrt{n} < \mu < \bar{X} + 1.64S/\sqrt{n}) = 0.9$$

$$=P[34223-1.64(2126.058) < \mu < 34223+1.64(2126.058)] = 0.9$$

$$=P[30736 < \mu < 37709] = 0.9$$

$$P(\bar{X} - 1.96S/\sqrt{n} < \mu < \bar{X} + 1.96S/\sqrt{n}) = 0.95$$

$$=P[34223-1.96(2126.058) < \mu < 34223+1.96(2126.058)] = 0.95$$

$$=P[30055 < \mu < 38390] = 0.95$$

The confidence intervals are wider for higher levels of certainty (e.g., 99% vs. 95%). That is, the wider the range, the more certain we can be that it will encompass the population mean. The CIs for categories with smaller sample sizes are wider because the SE is higher.

**4.** For the sample of householders who use non-public transit methods, personal wage earnings averaged \$39,961 per year with a standard deviation of \$47,161. For householders using public transit modes, earnings averaged \$34,223 per year with a standard deviation of \$24,975. The average income was significantly lower for householders using public transit modes. Thus, even after accounting for the fact that I have only a random sample, I could rule out most chances that income for the 2 groups were the same.

**5.** This memorandum responds to your request for information on the influence of personal wage earnings on decisions to use public transportation for commuting. Using Washington state data from the 2002 Population Survey, I found that average wage earnings were lower for residents who used public transportation to commute to work (bus and ferry) than for those who did not use public transportation modes. On average, earners using public transit made between \$30,055 and \$38,390 per year.<sup>1</sup> Earnings for non-public transit users were between \$38,225 and \$41,666 per year. These results suggest that householders with higher wage earnings are less likely to choose public transportation for their commuting preferences. Further research on demographic characteristics of households and decision making about the mode of commuting could allow you to develop policies that will increase the use of public transportation.

---

<sup>1</sup>This and all other range estimates are 95% confidence intervals.