Problem Set 9 - Solutions

Statistics - NYU, Spring 2016

Ercan Karadas

[1] X: score of test. $X \sim N(70, 8^2)$

a) $P(X \leq a) = 0.85$

$P(\frac{X-70}{8} \leq \frac{a-\mu}{\sigma})$=0.85

$\frac{X-70}{8} = 1.036$

$X = 78.288$

b) $P(X \leq a) = 0.22$

$P(\frac{X-70}{8} \leq \frac{a-\mu}{\sigma})$=0.22

$\frac{X-70}{8} = -0.7722$

$X = 63.82$

[2] X : Undergraduate GPA; $X \sim N(?, \quad .45^2)$; $n = 25$; $\bar{X} = 2.90$

a) $(1-\alpha)100 = 95 \Rightarrow \alpha = 0.05$
95% confidence interval for the population mean (μ):
$\bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ and $z_{\alpha/2} = z_{0.025} = 1.96$
$2.9 \pm 1.96\frac{0.45}{\sqrt{25}} \Rightarrow [2.7236, 3.0764]$

Interpretation: If we select samples of size 25, and calculate 95% confidence interval for the population mean (μ) repeatedly, as we did above for one such sample, 95% of these intervals would contain the true population mean (μ).
Remark: This does NOT mean that $P(2.7236 \leq \mu \leq 3.0764) = 0.95$. Because true population mean (μ) is a constant number (though unknown at that point), and therefore this specific interval contains the true population mean or not.

b) If everything else is the same, a higher confidence level results in a wider interval than as found in part (a).

c) Higher standard deviation means dispersion of the random variable is higher, which in turn result in a wider confidence interval than found in part (a). You could see this just by looking at how changing the standard deviation effects the confidence interval in the formula of C.I. without doing any calculations.

d) Increasing the sample size reduces the variance of the sample means ($\sigma_{\bar{X}^2} = \frac{\sigma^2}{n}$), which in turn means $\bar{X}$ becomes a better estimator of the true population mean (μ), and therefore the confidence interval will be narrower than the one found in part (a). You can see this just looking at how changing n effects the confidence interval in the formula of C.I. without doing any calculations.

e) $2.99 - 2.90 = ME = z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$
$0.09 = z_{\alpha/2}\frac{0.45}{\sqrt{25}}$
Solving this equation yields $z_{\alpha/2} = 1$
Then $\alpha = 2[1 - F(1)] = 0.3174$, and therefore confidence level is $100(1 - .3174)\% = 68.26\%$.

f) The only difference from part (a) is that here we do not know the population variance, so we need to use t-distribution instead of z-distribution.
95% confidence interval for the population mean (μ):
$\bar{X} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$ and $t_{\alpha/2,n-1} = t_{0.025,24} = 2.064$

95% C.I. : $2.9 \pm 2.064\frac{0.40}{\sqrt{25}} \Rightarrow [2.7349, 3.0651]$

[3] a) The sample data

$$n = 44, \quad \overline{X} = 126, \quad s^2 = 15$$

and since the true population variance is unknown we use $t-$distribution to construct the C.I. for μ:

$$\overline{X} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

$(1 - \alpha)100\% = 90\% \Rightarrow \alpha = 0.10$ and $n - 1 = 43$, and we find $t_{n-1,\alpha/2}$ from the t-table as $t_{43,0.05} = 1.684$. Then 90% C.I. for μ:

$$126 \mp 1.684\frac{15}{\sqrt{44}} = \big[122.2, \quad 129.8\big]$$

Remark: Note that t-table does not have exactly $t_{43,0.05}$ so I pick the closest one, which is $t_{40,0.05}$.

b) Interpretation is the same as before: If we select samples of size 44, and construct 90% confidence intervals for μ for each sample, as we did above for one such sample, 90% of these intervals will contain the true population mean, μ.

c) Since the consultant's prediction lies outside the C.I. that we have just calculated, our analysis, based on the available data, does not support the consultant.

[4] a) We would like to find a and b satisfying the following

$$P(a < \sigma^2 < b) = 1 - \alpha$$

If we had such (r.v.) a and b that are based on the sample data then we would be able to say $100(1-\alpha)\%$ of the time the constructed interval would catch the true population variance.

b)

$$\begin{aligned} a < \sigma^2 < b = \frac{1}{b} < \frac{1}{\sigma^2} < \frac{1}{a} \\ = \frac{(n-1)s^2}{b} < \frac{(n-1)s^2}{\sigma^2} < \frac{(n-1)s^2}{a} \end{aligned}$$

so we have

$$P\big(\frac{(n-1)s^2}{b} < \frac{(n-1)s^2}{\sigma^2} < \frac{(n-1)s^2}{a}\big) = 1-\alpha$$

c) In the middle we have the Chi-square distribution with $(n-1)$ degrees of freedom, $\chi^2_{n-1} = \frac{(n-1)s^2}{\sigma^2}$. Assuming we would like leave $\alpha/2$ probability in both tails we should have

$$P\big(\chi^2_{n-1,1-\frac{\alpha}{2}} < \chi^2_{n-1} < \chi^2_{n-1,\frac{\alpha}{2}}\big) = 1-\alpha$$

Comparing these last two equations, we should have

$$\frac{(n-1)s^2}{b} = \chi^2_{n-1,1-\frac{\alpha}{2}} \quad \text{and} \quad \frac{(n-1)s^2}{a} = \chi^2_{n-1,\frac{\alpha}{2}}$$

Solving for a and b we conclude

$$a = LCL = \frac{(n-1)s^2}{\chi^2_{n-1,\frac{\alpha}{2}}} \quad \text{and} \quad b = UCL = \frac{(n-1)s^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}$$

You will not be tested for this kind of derivations, but I think, it is worth knowing how the main formulas can be derived. More importantly, being familiar with that kind of manipulations helps a lot when we solve some other problems that require similar tricks. For example in the last problem in this problem set.

[5] a) Let X be the fat content of yogurt per cup, so $\overline{X}$ average fat content per cup of a sample of size 400, and we would like to find $P(\overline{X} > 0.52) = ?$.

Since the population variance σ^2 is unknown, student's t distribution should be used to compute the probability:

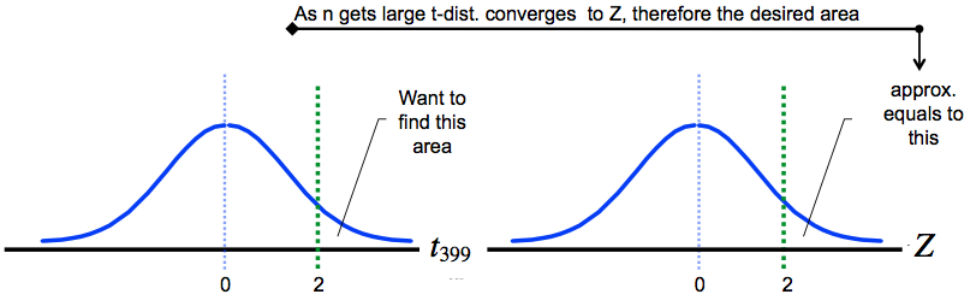
$$\begin{aligned} P(\overline{X} > 0.52) &= P\big(\frac{\overline{X}-\mu}{s/\sqrt{n}} > \frac{0.52-\mu}{s/\sqrt{n}}\big) \\ &= P\big(t_{n-1} > \frac{0.52-\mu}{s/\sqrt{n}}\big) \end{aligned}$$

So far everything is the same as we did with Z distribution, except here since σ is unknown instead of writing $Z = \frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$, we write $t_{n-1} = \frac{\overline{X}-\mu}{s/\sqrt{n}}$ on the right hand side of the inequality. Then, substitute what we know to get

$$\begin{aligned} P(\overline{X} > 0.52) &= P\big(t_{n-1} > \frac{0.52-\mu}{s/\sqrt{n}}\big) \\ &= P\big(t_{399} > \frac{0.52-0.50}{0.2/\sqrt{400}}\big) \\ &= P\big(t_{399} > 2\big) \end{aligned}$$

This is the area to the left of 2 for the random variable t_{399}, see the figure on the left below. But since $n = 400$ is large we know that $t_{n-1} \longrightarrow Z$, therefore we can substitute $P\big(t_{399} > 2\big)$ by $P(Z > 2)$ (see the figure on the right):

$$\begin{aligned} P(\overline{X} > 0.52) &= P\big(t_{399} > 2\big) \\ &= P(Z > 2) \\ &= 1 - 0.977 \\ &= 0.023 \end{aligned}$$



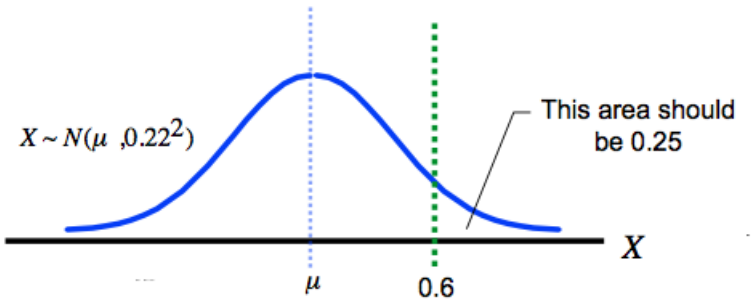
Remark: When n is large ($n > 100$), $t-$distribution is very close to the Z distribution, that is why after $n = 100$, it jumps to t_∞ in the t-table in your book. These, t_∞ values are nothing but the corresponding Z-values. For example, check that $t_{\infty,\,0.05} = Z_{0.05}$.

b) The 95% C.I. :

$$\overline{X} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 0.52 \mp t_{399,0.025}\frac{0.2}{\sqrt{400}} = [0.50, 0.54]$$

The company advertisement claims that $\mu = 0.50$, but our sample data indicate that it is very unlikely that the claim is true for $\mu = 50$ falls outside the confidence interval.

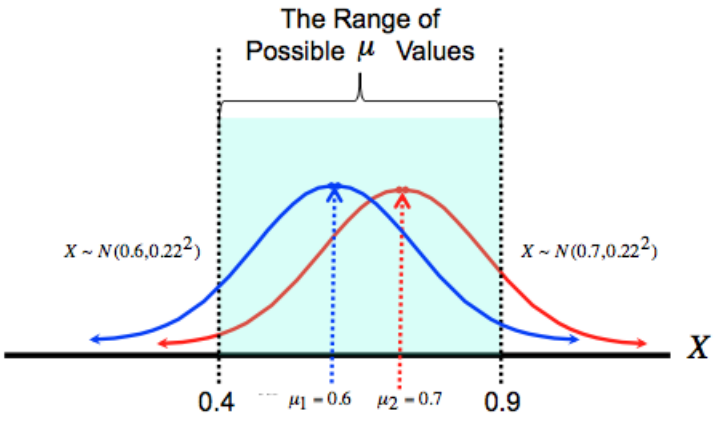
c) We know that $X \sim N(\mu, 0.22^2)$, but do not know the μ yet, i.e. we know that the distribution is normal with $\sigma = .22$, but we still need to determine its location from the given information. We can do that by finding a μ such that indeed $P(X > 0.6) = 0.25$ as claimed by the company (see the figure below):



$$\begin{aligned} P(X > 0.6) &= P(Z > \frac{0.6 - \mu}{0.22}) = 0.25 \\ &\Rightarrow \frac{0.6 - \mu}{0.22} = 0.675 \quad (= Z_{0.25}) \\ &\Rightarrow \mu = 0.45 \end{aligned}$$

d) The following could be an example for the first statement: the fat content of the yogurt produced is determined by the overall quality of the production facilities used, say this is the type of the company, which we denote by μ here. Furthermore, this overall quality is uniformly distributed on $[0.4, 0.9]$. It is as if when the company first established its type (μ) is determined by a random draw from [0.4, 0.9] interval, and then the fat content is normally distributed with the mean equal to the realized type.

In the following figure, the green area indicates all possible initial overall facility qualities (types). It also depicts what the distribution of the fat content would look like if the initial quality draw were $\mu_1 = 0.6$ or $\mu_2 = 0.7$.



We need to determine an initial type draw, say μ^*, so that the distribution of the fat content is $N(\mu^*, 0.22^2)$ and the area to the right of 1.1 is 0.06 (i.e. at least 24 cups, out of 400, will contain more than 1.1 units of fat):

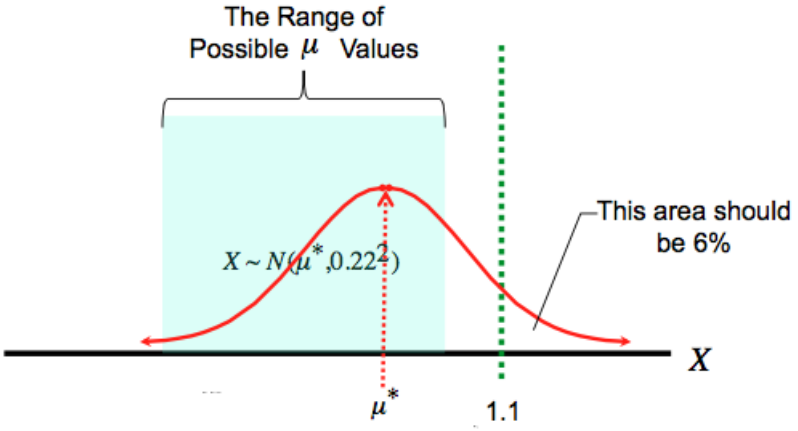
$$\begin{aligned} P(X > 1.1) &= P(Z > \frac{1.1 - \mu^*}{0.22}) \\ &\Rightarrow \frac{1.1 - \mu^*}{0.22} = 1.56 \quad (= Z_{0.06}) \\ &\Rightarrow \mu^* = 0.76 \end{aligned}$$

Therefore, if the initial draw were $\mu^* = 0.76$ or higher indeed 6% of the total production would contain more than 1.1 units of fat.

Finally, since μ is uniformly distributed on [0.4, 0.9],

$$P(\mu > \mu^*) = P(\mu > 0.76) = \frac{0.9 - 0.76}{0.9 - 0.4} = \frac{0.14}{0.50} = 0.28$$

is the probability that at least 6% of the yogurt produced will contain more than 1.1 units fat.



[6]

Expected # of accidents $= P(\sigma < 32)(1.5) + P(32 < \sigma < 60)(2.8) + P(\sigma > 60)(4.5)$

Let us first compute the probabilities:

$$\begin{aligned}
P(\sigma < 32) =& P(\frac{1}{\sigma} > \frac{1}{32}) \\
=& P(\frac{s}{\sigma} > \frac{s}{32}) = P(\frac{s^2}{\sigma^2} > \frac{s^2}{32^2}) \\
=& P(\frac{(n-1)s^2}{\sigma^2} > \frac{(n-1)s^2}{32^2}) \\
=& P(\chi^2_{(n-1)} > \frac{(n-1)s^2}{32^2}) \\
=& P(\chi^2_{18} > \frac{(18)40^2}{32^2}) \\
=& P(\chi^2_{18} > 28.125) \\
=& 0.05
\end{aligned}$$

Similarly,

$$\begin{aligned}
P(\sigma < 60) =& P(\frac{1}{\sigma} > \frac{1}{60}) \\
=& P(\chi^2_{18} > \frac{(18)40^2}{60^2}) \\
=& P(\chi^2_{18} > 8) \\
=& 0.975
\end{aligned}$$

Therefore,
$P(32 < \sigma < 60) = P(\sigma < 60) - P(\sigma < 32) = 0.975 - 0.05 = 0.925$
and

$P(\sigma > 60) = 1 - P(\sigma < 60) = 1 - 0.975 = 0.025$
Finally, putting all these together we find the answer as

$$\begin{aligned}\text{Expected \# of accidents} =& P(\sigma < 32)(1.5) + P(32 < \sigma < 60)(2.8) + P(\sigma > 60)(4.5)\\ =& (0.05)(1.5) + (0.925)(2.8) + (0.025)(4.5)\\ =& 0.075 + 2.59 + 0.1125\\ =& 2.78\end{aligned}$$

[7] X: amount of coke in each bottle.

a) LCL: $a = \bar{X} - z_{0.015}\frac{1.2}{\sqrt{20}} = 35 - 2.170\frac{1.2}{\sqrt{20}} = 34.418$

UCL: $b = \bar{X} + z_{0.015}\frac{1.2}{\sqrt{20}} = 35 + 2.170\frac{1.2}{\sqrt{20}} = 35.582$

These are respectively the lower and upper bounds of the 97% confidence interval with sample size 20.

b) The process in not in control for any of the samples.

[8] a) We want to construct a $100(1-\alpha)\%$ confidence interval for μ, i.e. an interval such that μ lies in the interval $1-\alpha$. Hence a and b such that $P(a \le \mu \le b) = 1 - \alpha$ are the bounds of the confidence interval we are looking for. Note that here a and b are random variables.

b)

$$\begin{aligned}P(a < \mu < b) &= 1 - \alpha\\ P(-b < -\mu < -a) &= 1 - \alpha\\ P(\overline{X} - b < \overline{X} - \mu < \overline{X} - a) &= 1 - \alpha\\ P(\frac{\overline{X} - b}{\sigma/\sqrt{n}} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{\overline{X} - a}{\sigma/\sqrt{n}}) &= 1 - \alpha\end{aligned}$$

c) Since $\frac{\overline{X}-a}{\sigma/\sqrt{n}} = z_{\alpha/2}$, we can solve for a to obtain $a = \overline{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.
Similarly, since $\frac{\overline{X}-b}{\sigma/\sqrt{n}} = -z_{\alpha/2}$, we can solve for b to obtain $b = \overline{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

[9] a)

$$n = 15, s = 2.36, \text{ and } \alpha = 0.05$$

C.I. for population variance:

$$\begin{aligned}\Big[\frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}, \quad \frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}\Big] &= \Big[\frac{14(2.36)^2}{\chi^2_{14,0.025}}, \quad \frac{(14)(2.36)^2}{\chi^2_{14,0.975}}\Big]\\ &= \Big[\frac{14(2.36)^2}{26.119}, \quad \frac{(14)(2.36)^2}{5.269}\Big]\\ &= \big[2.98, \quad 13.85\big]\end{aligned}$$

b) Without any new calculation We can say that it should be wider. Because in order to be able to say that the constructed C.I. contains the true variance more often width of the interval should be increased.

[10] a) Let X denote the arrival delay (so $X = 0$ means it is on time, $X < 0$ indicates arrivals before 7.35, and $X > 0$ after 7.35). We are given

$$n = 200, \quad \overline{X} = 1, \text{ and } \quad s = 2.3$$

We also know that X is normally distributed, but neither the mean nor the standard deviation are known, so let us denote this distribution as $X \sim N(\mu, \sigma^2)$, where both μ and σ^2 are unknown at the moment.

Since the sample average $\overline{X}$ is an unbiased estimator for the true population mean, μ, our best (point) estimate for the average delay $E(X)$ is

$$\overline{X} = 1$$

which in return implies that based on the sample data our estimate for the average arrival time is 7.36 A.M.

b) The 90% C.I. for the average arrival time, μ, (in minutes) :

$$\overline{X} \mp t_{n-1,\alpha/2}\frac{s}{\sqrt{n}} = 1 \mp 1.645\frac{2.3}{\sqrt{199}} = 1 \mp 0.27$$

In terms of the clock time, the confidence interval will be

$$7.36 \; A.M. \mp 0.27 \text{ min.}$$

(Note that, again, since n is large, both $t_{199,0.05} = t_{\infty,0.05}$ and $t_{199,0.05} = Z_{0.05}$ are correct).

c) You would miss the train only if $X < -1$, i.e., it arrives at before 7.34, and the probability that this happens can be computed as:

$$\begin{aligned} P(X < -1) &= P(Z < \frac{-1-0}{\sqrt{1.23}}) \\ &= P(Z < -0.9) \\ &= 1 - 0.8159 = 0.1841 \end{aligned}$$

Therefore, you should expect to miss $200 \times (0.1841) = 37$ trains.

[11] $X \sim N(70, 8^2)$

a) $1 - 2 \times (1 - P(X \leq 1)) = 0.6826$

b) $2 \times (1 - P(X \leq 2)) = 0.0456$

c) $P(X \leq 2.1) - P(X \leq 1.25) = 0.9821 - 0.8944 = 0.0877$