Statistics

Solutions to Practice Questions

ESD, SUTD

Term 5, 2017

Q1. (a) $\binom{10^6-100}{1000}/\binom{10^6}{1000}$.

(b) $(1-100/10^6)^{1000} = (1-10^{-4})^{1000}$.

(c) Yes, since the total number of items is very large, sampling with or without replacement should roughly give the same result.

(d) Recall that from either the Taylor or the binomial expansion, $(1+x)^n \approx 1+nx$ when $|x|$ is small.

From part (b), we have $x = -10^{-4}$ and $n = 1000$, so the answer is roughly $1 - 1/10 = 0.9$.

Q2. Engine size: numerical, continuous.

Number of cylinders: numerical, discrete.

Size of car: categorical, ordinal.

Type of transmission: categorical, nominal.

Dealer cost: numerical, discrete (or, essentially continuous).

Q3. There are 100 male managers, 50 female managers, 1900 male employees, and 1450 female employees.

Because the sample size relative to the population size is $700/3500 = 1/5$, therefore we randomly sample $1/5$ of the people from each of these strata, namely:

20 male managers, 10 female managers, 380 male employees, 290 female employees.

Q4. No, not every 40-buyer sample has an equal chance of being chosen.

In a simple random sample, 40 buyers are chosen randomly from all 1000 buyers, so most of the time we will not get 10 buyers from each brand.

The method used here is actually stratified sampling.

Q5. $\bar{x} = 1 + x/3$.

The sample variance $= 1^2 =$
$\frac{1}{2}\big((1-\bar{x})^2 + (2-\bar{x})^2 + (x-\bar{x})^2\big) = 1 - x + x^2/3.$

Therefore $x = x^2/3$, so $x = 3$ or $x = 0$.

Note that it is easy to check your answers for this question.

Q6. An intuitive method is to 'scale' one group so it has the same size as the other, and extrapolate the total. For instance, if we scale the first group so it ends up with size n, then the corresponding total would be $u_1 \frac{n}{m}$. Thus the required proportion (using the method taught in class) is

$$\left(u_2 - u_1 \frac{n}{m}\right) \frac{1}{n} = \frac{u_2}{n} - \frac{u_1}{m}.$$

Scaling the second group gives the same answer.

A more rigorous approach is as follows. Let X denote the random variable for the number of upsetting items from the first three questions, and let Y represent the number of upsetting items (either 0 or 1) from the fourth (sensitive) question. Then $\frac{u_1}{m}$ is an estimator for $\mathsf{E}(X)$, and $\frac{u_2}{n}$ is an estimator for $\mathsf{E}(X+Y)$. Thus $P(Y) = \mathsf{E}(Y) = \mathsf{E}(X+Y) - \mathsf{E}(X)$ can be estimated by $\frac{u_2}{n} - \frac{u_1}{m}$.

Q7. We cannot always reject at the 1% level, since the p-value could be between 1% and 5%.

However, we can always reject at the 10% level, since the p-value is less than 5%.

Q8. (a) There could be confounding variables that explain both eye problems and employability, such as how much one studied (or read, etc) at university.

(b) Again, there could be confounding variables, for instance, it might be the case that people from higher socioeconomic background live longer, and are also more likely to be vegetarian.

(c) Even if there was only 1 smoker 35 years ago, the current number of smokers would be roughly $2^{35} = 2^5 \times (2^{10})^3 \approx 32$ billion, which is greater than the world's population.

Q9. No, 4 appears twice in the 2nd column, also 3 appears twice in the 3rd column.

Q10. No. Even though $2\bar{x} - 1$ is inspired by the German tank problem, the situation here is different in a number of ways. Firstly, the number of customers who order from the shop may not be fixed, but changes from day to day.

Moreover, the visits from the other customers are probably not uniformly distributed, but clustered around peak hours (such as lunch and dinner times). As an extreme example of this, suppose all other customers visit the shop just before closing time. Then, there is a very high probability that you observe $x_i = 1$, and so $2\bar{x} - 1$ is around 1 and gives no information about the total number of customers.

Q11. Roughly 50% of the data values are greater than 140, and 25% of the data values are greater than 180. So (roughly) between 25% and 50% of the data values are greater than 160.

Q12. 25% of all values are higher than Q_3, so Q_3 should be $\Phi^{-1}(0.75) = 0.675$ sd above the mean, and Q_1 should be 0.675 sd below the mean.

Therefore, to be 1.5 IQR below Q_1 translates to 2.7 sd below the mean, and 1.5 IQR above Q_3 to 2.7 sd above the mean. So the probability of being classified as an outlier is about $2P(Z > 2.7) = 0.007$.

Q13. The quickest way is to swap the roles of μ and μ_0 in the formula given on the Week 5 slide, 'Power calculation – formula', since that formula is for $H_1 : \mu > \mu_0$. The result is

$$1 - \beta = \Phi\left(\frac{(\mu_0 - \mu)\sqrt{n}}{\sigma} - z_{1-\alpha}\right).$$

Q14. Let the numbers of successes be denoted by a, b, c, d:

	Male	Female
Treatment A	$a/50$	$b/90$
Treatment B	$c/60$	$d/30$

From the information given, we have $a + b = 70$, $c + d > 45$.

For Simpson's paradox to occur, we need $a/50 > c/60$ and $b/90 > d/30$.

Any numbers that satisfy the above four relations will work, for instance, we can take $a = c = 50$, $b = 20$, $d = 0$.

Q15. (a) Sort the male temperatures and the female temperatures separately, then use a scatter plot.

(b) Simple example: testing whether a certain medication affects body temperature; each pair consists of the before and after temperatures from the same person.

Q16. (a)

(Note that (c) is nonsense, since we never accept the null.)

Q17. β cannot be determined in either case, since β is not only a function of α and n, but also of σ and $\mu - \mu_0$. (See for instance the answer to Q13.)

Q18. The width of the confidence interval is $2z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. To halve the width, n needs to be increased by a factor of 4.

Q19. EWMA uses all the previous values, as can be seen from recursively applying the formula
$\text{EWMA}_t = \alpha\, x_t + (1-\alpha)\, \text{EWMA}_{t-1}$.

Ordinary moving average only uses a fixed number (not all) of the previous values.

Q20. The 90% confidence interval contains the 80% CI, so its lower limit is less than 7.5 and its upper limit is greater than 8.3. Moreover, it should be symmetric around $\bar{x} = (7.5 + 8.3)/2 = 7.9$. (d) is the only one the satisfies these constraints.

Q21. (a) is wrong because it does not contain $s^2 = 5$. (c) is wrong because it is a one-sided confidence interval. (d) is wrong because variances can never be negative.

(e) is also wrong because it is symmetric around 5, but CI's for variance are not symmetric. This leaves (b) as the only option.

Q22. There are $n = 52$ weeks in a year. By the CLT, the average is approximately normally distributed with mean 312 and sd $58/\sqrt{52}$.

Therefore the required probability is approximately
$P(\bar{X} > 330) = P(Z > (330 - 312)/(58/\sqrt{52})) = 0.0126$.

Q23. For this, we may use the formula given in Homework 2 Question 5. Here, $|\mu - \mu_0| = 2$, $n = 5$, $\sigma = 2$, $\alpha = 0.01$. Then

$$\begin{aligned} 1 - \beta &= \Phi\left(\frac{(\mu - \mu_0)\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) + \Phi\left(\frac{(\mu_0 - \mu)\sqrt{n}}{\sigma} - z_{1-\alpha/2}\right) \\ &= \Phi\left(\frac{2\sqrt{5}}{2} - z_{0.995}\right) + \Phi\left(\frac{-2\sqrt{5}}{2} - z_{0.995}\right) = 0.367. \end{aligned}$$

Similarly, if $|\mu - \mu_0| = 4$, then the power is 0.971.

Q24. (a) $\bar{x} = 200$, $\alpha = 0.1$.

Two-sided 90% CI: $[199.984,\ 200.0164]$;

upper 90% CI: $(-\infty,\ 200.0128]$ (or, effectively, $[0,\ 200.0128]$);

lower 90% CI: $[199.987,\ \infty)$.

(b) $s = 0.02708$. The 90% two-sided CI for σ is $[0.017,\ 0.079]$, which contains $\sigma = 0.02$, so s is consistent with the given σ.

Q25. Since the sample size $n = 196$ is quite large, CLT applies, and also σ can be well approximated by s, therefore the z-distribution is a good approximation.

The confidence interval is approximately

$$\left[\bar{x} - 1.96\frac{s}{\sqrt{n}},\ \bar{x} + 1.96\frac{s}{\sqrt{n}}\right] = [17.104,\ 18.896].$$

Q26. Due to the normality assumption, the sample mean is also normally distributed.

We compute $z = (117.5 - 115)/(25/\sqrt{25}) = 0.5$. The p-value is one-sided, given by $P(Z > z) = 0.3085$.

Q27. $H_0 : \mu = 16.2$, $H_1 : \mu < 16.2$.

Normality and known σ together imply that we can use the formula in the answer to Q13. Here $\mu_0 = 16.2$, $\mu = 14.6$, $\sigma = 2.4$, $n = 16$ and $\alpha = 0.05$.

Plugging these values into the formula, we find that the power is $\Phi(1.0218) \approx 0.847$.

Q28. $P(s^2 > 2\sigma^2) = P(4s^2/\sigma^2 > 8) = P(\chi^2_4 > 8)$, since $(n-1)s^2/\sigma^2$ is a χ^2_{n-1} random variable.

The pdf of a χ^2_4 random variable is $\frac{1}{4}xe^{-x/2}$.

Therefore, the required probability is

$$\int_8^\infty \frac{1}{4}xe^{-x/2}\,\mathrm{d}x = \left[-e^{-x/2} - \frac{x}{2}e^{-x/2}\right]_8^\infty = \frac{5}{e^4} \approx 0.09158,$$

where the integration was done by parts.