Tentamen 16 december 2014, antwoorden

Statistical Methods (Vrije Universiteit Amsterdam)

Solutions Exam Empirical Methods

VU University Amsterdam, Faculty of Exact Sciences

15.15 – 18.00h, December 16, 2014

1. (a) True. With a Pareto bar chart it is directly visible which category has the largest frequency count, even for small differences with other categories.
   (b) False. In a stratified sample, the population is divided according to different strata and a random sample is taken from each stratum.
   (c) False. Zero degrees Celsius is not a natural zero point, so outside temperatures are at interval level of measurement.
   (d) False. Consider the dataset 1,2,6. The median is 2, which is smaller than the mean 3.
   (Other examples/arguments are of course also possible)

2. (a) Note that $P(\text{warm}) = 1 - P(\text{cold}) - P(\text{mild}) = 1 - 0.30 - 0.45 = 0.25$. Using the law of total probability we get

$$\begin{aligned} P(\text{rain}) &= P(\text{rain}|\text{cold}) \cdot P(\text{cold}) + P(\text{rain}|\text{mild}) \cdot P(\text{mild}) + P(\text{rain}|\text{warm}) \cdot P(\text{warm}) \\ &= 0.30 \cdot 0.30 + 0.10 \cdot 0.45 + 0.05 \cdot 0.25 = 0.1475. \end{aligned}$$

   (b) A and B are independent events if $P(A) \cdot P(B) = P(A \cap B)$. Clearly, $P(A) = 0.45$. By the complement rule, $P(B) = 1 - P(\text{rain}) = 1 - 0.1475 = 0.8525$, so $P(A) \cdot P(B) = 0.45 \cdot 0.8525 \approx 0.384$. By the multiplication rule,

$$P(A \cap B) = P(B|A) \cdot P(A) = 0.90 \cdot 0.45 = 0.405.$$

   So A and B are not independent.

   (c) According to the addition rule,

$$P(\text{cold or rain}) = P(\text{cold}) + P(\text{rain}) - P(\text{cold and rain}).$$

   Again by the multiplication rule,

$$P(\text{cold and rain}) = P(\text{rain}|\text{cold}) \cdot P(\text{cold}) = 0.30 \cdot 0.30 = 0.09.$$

   Hence, $P(\text{cold or rain}) = 0.30 + 0.1475 - 0.09 = 0.3575$.

   (d) By definition of conditional probability,

$$P(\text{warm}|\text{rain}) = \frac{P(\text{warm and rain})}{P(\text{rain})}.$$

   Using the multiplication rule again and the answer of part a), we get

$$P(\text{warm}|\text{rain}) = \frac{P(\text{rain}|\text{warm}) \cdot P(\text{warm})}{P(\text{rain})} = \frac{0.05 \cdot 0.25}{0.1475} \approx 0.085.$$

   The same answer can be obtained by using Bayes' Theorem.

3. (a) Weight of single pack is normally distributed with mean $\mu = 1.01$ and $\sigma = 0.012$, so the z score of $x = 1.00$ is $z = \frac{1.00-1.01}{0.012} \approx -0.83$. Looking up this value in Table 2 yields that the required probability equals 0.2033.

   (b) The weight of $n = 16$ sugar packs is normally distruted with mean $\mu = 1.01$ and $\sigma = 0.012/\sqrt{16} = 0.03$. So the the z score of $x = 1.00$ is now $z = \frac{1.00-1.01}{0.03} \approx -3.33$. Looking up this value and using that 'area to the right = 1- area to the left' the required probability is 1-0.0004 = 0.9996.

   (c) Since σ is unknown the general formula for a $1-\alpha$ confidence interval is given by $\overline{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}}$. Since $n = 25$ and $\alpha = 0.10$ we have $t_{24,0.05} = 1.711$. Together with the sample statistics this yields the following 90% CI:

$$\overline{x} \pm t_{n-1,\alpha/2} \cdot \frac{s}{\sqrt{n}} = 1.005 \pm 1.711 \cdot \frac{0.008}{\sqrt{25}} = 1.005 \pm 0.003 = [1.002, 1.008]$$

   (d) If we take many samples of size $n = 25$ and construct a 90% CI for each sample, then on average 90% of these intervals would contain the true unknown population parameter μ.

   (e) No, since 1.01 is not contained in the CI it seems unreasonable that $\mu = 1.01$ with a 90% confidence level.

4. (a) The probability p that the random-number generator produces a 0. Point estimate $\hat{p} = \frac{25{,}264}{50{,}000} \approx 0.505$.

   (b) We follow the steps of hypothesis testing:

      1. $H_0 : p = 0.50$;
         $H_1 : p \neq 0.50$.
         Significance level $\alpha = 0.01$.
      2. Test statistic: $Z = \frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$ has under H_0 approximately a $N(0,1)$-distribution.
         NB: $n = 50,000 > 30$ so requirement for approximation is met.
      3. Observed score:
$$z = \frac{0.505 - 0.50}{\sqrt{0.5 * 0.5/50,000}} \approx 2.24.$$
         (Other values possible by other rounding.)
      4. Since the test is two-tailed, we have
$$\begin{aligned} P - \text{value} &= 2\min\{P(Z \geq 2.24), P(Z \leq 2.24)\} \\ &= 2P(Z \geq 2.24) \approx 2 * (1 - 0.9875) = 0.013. \end{aligned}$$
      5. Since P-value $= 0.013 > 0.01 = \alpha$ we fail to reject H_0.
      6. There is not sufficient evidence to warrant rejection of the claim that 0s and 1s occur with equal probability.

   (c) The formula to determine the required sample size in this case is $n = \left(\frac{z_{\alpha/2}}{2E}\right)^2$. Since $\alpha = 0.01$ and $E = 0.002$ we get $n = \left(\frac{2.575}{2 \cdot 0.002}\right)^2 = 414,414.1$ So 414,415 values should be investigated.

5. (a) Since both samples are independent, the population standard deviations are unknown and there is no reason why $\sigma_1 = \sigma_2$, we carry out the two-sample t-test for independent samples assuming that $\sigma_1 \neq \sigma_2$:

   1. $H_0 : \mu_1 = \mu_2$, where μ_1 is population mean of the test scores of group 1, and μ_2 of group 2.
      $H_1 : \mu_1 < \mu_2$.
      Significance level $\alpha = 0.05$.
   2. Test statistic: $T_2 = \frac{(\bar{x}_1-\bar{x}_2)-(\mu_1-\mu_2)}{\sqrt{s_1^2/n_1+s_2^2/n_2}}$ has under H_0 a t-distribution with approximately $\tilde{n}$ degrees of freedom, where $\tilde{n} = \min\{n_1 - 1, n_2 - 1\}$.
   3. Observed value:
   $$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{6.13 - 6.66 - 0}{\sqrt{1.12^2/91 + 1.91^2/91}} \approx -2.28.$$
   4. Critical value(s): test is left-tailed, $\alpha = 0.05$ and $\tilde{n} = 90$, so the critical value is $-t_{90,0.05} = -1.662$.
   5. Since $t = -2.28 < -1.662$ we reject H_0.
   6. There is sufficient evidence to support the claim that the new teaching method is better than the standard method.

   (b) Either both samples should be from a normally distributed population or both $n_1 > 30$ and $n_2 > 30$. Since $n_1 = 91 > 30$ and $n_2 = 91 > 30$, the latter holds.

6. (a) Test of homogeneity, we view the three drugs as three different populations and look whether they cause the same proportion of allergic reactions.

   (b) 1. H_0: drugs A, B, C have same proportion of allergic reactions;
      H_1: drugs A, B, C do not have the same proportion of allergic reactions.
      Significance level $\alpha = 0.01$.
   2. Test statistic $X^2 = \sum \frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ has under H_0 approximately a chi-square distribution with $(r-1)(c-1) = (3-1)(2-1) = 2$ degrees of freedom.
   3. Observed value is $\chi^2 = 4.10$ (given).
   4. Critical value: since $(r-1)(c-1) = 2$, $\alpha = 0.01$ and the chi-square test is right-tailed the critical value is $\chi^2_{(r-1)(c-1),\alpha} = \chi^2_{2,0.01} = 9.210$.
   5. Since $\chi^2 = 4.10 < 9.210 = \chi^2_{2,0.01}$ we fail to reject H_0.
   6. There is not sufficient evidence to warrant rejection of the claim that the three drugs have the same proportion of allergic reactions.

   (c) All expected frequencies E_{ij} should be larger than 1 and 80% should be larger than 5. Since $E_{ij} = (\text{row total}) \cdot (\text{column total})/(\text{grand total})$ we find $E_{11} = 100 \cdot 210/300 = 70$. Similar computations yield $E_{12} = 30, E_{21} = 70, E_{22} = 30, E_{31} = 70, E_{32} = 30$. All are larger than 5 so the requirements are met.

   (d) No, a directed alternative hypothesis could only be tested in a 2×2 contingency table.

7. (a) $\hat{y} = b_0 + b_1 x = 65.34 + 11.34x$. Predicted download time for a file of size 5.0 MB is therefore $65.34 + 11.34 \cdot 5.00 = 122.04$ ms.

   (b) $r^2 \approx 0.865$.

(c) 1. $H_0 : \beta_1 = 0$;
$H_1 : \beta_1 \neq 0$.
Significance level $\alpha = 0.05$.

2. Test statistic: $T_1 = \frac{b_1}{s_{b_1}}$ has under H_0 a t-distribution with $n-2$ degrees of freedom.
3. Observed value: $t = \frac{11.34}{1.64} \approx 6.91$.
4. Critical values: two-tailed test, $n-2 = 7$ and $\alpha = 0.05$ so the critical values are $-t_{7,0.025} = -2.36$ and $t_{7,.025} = 2.36$.
5. Since $t = 6.91 > 2.36$ we reject H_0.
6. There is sufficient evidence to warrant rejection of the claim that there is no linear relationship between the explanatory variable file size and response variable download time.

(d)
- The errors should come from a normal distribution. The residuals are estimates for the errors, so according to the normal Q-Q plot of the residuals, which is approximately a straight line, it is reasonable to assume that this requirement is met.
- The standard deviation should be fixed. This can be checked with a residual plot: since there is no pattern or 'fan'-shape, it is reasonable to assume that this requirement is also met.

(e) The scatterplot is approximately a straight line, the test rejects no linear relationship and the requirements for the test are met, so the linear regression model seems an appropriate model for the data.