Statistics for Astronomers
Solutions to Homework #4

Prof. Sundar Srinivasan

November 6, 2020

1. (a) Since the X_i are identical and normally distributed, Y is a sum of squares of standardised normal variables. Therefore, it has a χ^2 distribution with N degrees of freedom.

   Since $Y \sim \chi^2(N)$, $\mathbb{E}[Y] = N$, $\mathrm{Var}[Y] = 2N$.

   $Y = \dfrac{N}{\sigma^2}$ MSE, where MSE is the mean square error (MSE).

   (b) According to the Central Limit Theorem, $\overline{X} \sim \mathscr{N}(\mu, \sigma^2/N)$. U is then the square of a standardised normal variable. Therefore, $U \sim \chi^2(1), \mathbb{E}[U] = 1, \mathrm{Var}[U] = 2$.

   (c) Using the hint provided in the question and the results above,

$$\sum_{i=1}^{N}\Big(X_i-\mu\Big)^2 = \sum_{i=1}^{N}\Big(X_i-\overline{X}\Big)^2 + N\Big(\overline{X}-\mu\Big)^2 \implies \sigma^2\ Y = \sigma^2\ W + \sigma^2\ U \implies W = Y - U$$

   W is therefore a linear combination of Y and U. We can easily compute its expectation value and variance:

$$\mathbb{E}[W] = \mathbb{E}[Y] - \mathbb{E}[U] = N - 1 \tag{1}$$

   The variance is not as easy to calculate; Y and U are not independent because $\overline{X}$ is computed from the X_i and therefore $\overline{X}$ is not independent of X_i. This means that the variance of the sum $X_i + \overline{X}$ is not the sum of their variances (there is a third term that relates to their covariance). Instead, we will have to guess the distribution of W and then attempt to identify the variance of this distribution.

   If S^2 is the (Bessel-corrected or unbiased) sample variance, then we have

$$W = (N-1)\frac{S^2}{\sigma^2} \implies S^2 = \sigma^2\frac{W}{N-1} \tag{2}$$

   Since W also involves the sum of squares of normal deviates, we would expect that it is also χ^2-distributed like Y and U. This is indeed the case, as can be shown using Cochran's Theorem. The number of degrees of freedom for W is $N-1$. This is due to the fact that it involves

$Y \sim \chi^2(N)$ but also involve one constraint due to U, where the expectation value $\mathbb{E}[X]$ is estimated from the sample mean $\overline{X}$.

Since $W \sim \chi^2(N-1)$, its variance is $2(N-1)$.

(d) From Equations (1) and (2), $\mathbb{E}[S^2] = \sigma^2$ (consistent with S^2 being an unbiased estimator of the population variance by definition), and

$$\mathrm{Var}[S^2] = \left(\frac{\sigma^2}{N-1}\right)^2 \mathrm{Var}[W] = 2(N-1)\left(\frac{\sigma^2}{N-1}\right)^2 = 2\frac{\sigma^4}{N-1}$$

Thus, the sample standard deviation of a normally-distributed random variable is $\chi^2(N-1)$-distributed, with mean σ^2 and variance $2\dfrac{\sigma^4}{N-1}$.

2. The sample mean for the passenger masses is $8450/100 = 84.5$ kg. The standard deviation of this sample mean is $\sigma/\sqrt{100} = 1.5$ kg. The 95% CI is approximately the 2σ CI, which is then $[81.5\text{ kg}, 87.5\text{ kg}]$.

3. In the frequentist interpretation, if we generate many 95% CIs using the same procedure, then we will "trap" the true parameter value in 95% of the CIs. The code available here outputs:

```
Out of 100 CIs, 94 trap the true mean (0).
```

The same code run 1000 times using the `prob3multi` module returns a mean fraction of approximately 0.95. This fraction is consistent with the expectation that, if the procedure for constructing a 95% CI is repeated a large number of times, the true mean will be captured in 95% of those CIs.

4. (a) The sample mean and (unbiased) standard deviation of the metallicities of Sarah's stars are 1.657 dex and 0.858 dex respectively. The standard deviation of the sample mean is therefore 0.429 dex. Since this was computed using the sample mean, we must use the t distribution to compute the confidence interval.

For an observation X, we first "studentise" X: $T = \dfrac{X - \overline{X}}{S_{\overline{X}}} = \dfrac{X - 1.657}{0.429}$.

The 95% CI is such that $P(|T| < t_{\alpha/2}) = 1 - \alpha = 0.95$. We can solve for $t_{\alpha/2}$ using `scipy.stats.t.ppf`:

```
print (scipy.stats.t.ppf((1-0.95)/2, df = 3)) #will have a negative sign
```

results in $t_{\alpha/2} \approx 3.182$ standard deviations, which we can convert back to a constraint on the observation X:

$$|T| \equiv \frac{X - \overline{X}}{S_{\overline{X}}} \leq t = 3.182 \Longrightarrow \overline{X} - 3.182\ S_{\overline{X}} \leq X \leq \overline{X} + 3.182\ S_{\overline{X}} \tag{3}$$

Thus, the 95% CI on the true mean is $[0.292\text{ dex}, 3.023\text{ dex}]$ (since metallicities are allowed to be negative, this CI is not include unphysical values).

(b) The lower bound to the 95% CI computed in above is greater than 0, the mean computed by Cassagrande et al. This means that the value 0 is greater than 3.18 standard deviations away, or outside the interval that includes 95% of the possible observations. This means that the significance level is less than 5%. Sarah must definitely publish these results!