Solutions to Problems 4

Chuizheng Kong

November 20, 2024

4.1

Suppose $X_1, X_2, \cdots, X_n$ are IID random variables. Let $\overline{X}$ be their sample average and use $E(X)$ and $\text{Var}(X)$ to denote the expectation and variance of these variables (they both exist). Further X and Y are 2 random variables and the expectation and variance both exist for these variables.

4.1.1

Prove that $\text{Var}(\overline{X}) = \dfrac{\text{Var}(X)}{n}$.

Solution: Since the X_is are independent,

$$\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = n \cdot \text{Var}(X) \tag{4.1.1}$$

Given that $\overline{X} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} X_i$, we get:

$$\text{Var}\left(\overline{X}\right) = \text{Var}\left(\frac{\sum_{i=1}^{n} X_i}{n}\right) = \frac{1}{n^2}\text{Var}\left(\sum_{i=1}^{n} X_i\right) = \frac{n \cdot \text{Var}(X)}{n^2} = \frac{\text{Var}(X)}{n} \tag{4.1.2}$$

4.1.2

Define $S^2 = \dfrac{1}{n-1}\displaystyle\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$. Show that $E\left(S^2\right) = \text{Var}(X)$ thus S^2 is an unbiased estimate of $\text{Var}(X)$.

Solution: Using Equation 4.1.2, we first consider:

$$\begin{aligned}
E\left[\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2\right] &= E\left[\sum_{i=1}^{n}\left(X_i^2 - 2X_i\overline{X} + \overline{X}^2\right)\right] \\
&= E\left(\sum_{i=1}^{n} X_i^2 - 2n\overline{X}^2 + n\overline{X}^2\right) \\
&= E\left(\sum_{i=1}^{n} X_i^2\right) - nE\left(\overline{X}^2\right) \\
&= nE\left(X^2\right) - n\left[\text{Var}\left(\overline{X}\right) + E^2\left(\overline{X}\right)\right] \\
&= n\left[\text{Var}(X) + E^2(X)\right] - \text{Var}(X) - nE^2(X) \\
&= (n-1)\text{Var}(X)
\end{aligned} \tag{4.1.3}$$

Therefore,

$$
\begin{aligned}
E(S^2) &= E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2\right] \\
&= \frac{1}{n-1}(n-1)\mathrm{Var}(X) \\
&= \mathrm{Var}(X) \qquad (4.1.4)
\end{aligned}
$$

4.1.3

Show that $\mathrm{Cov}(X,Y) = E(XY) - E(X)E(Y)$ and $\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y)$.

Solution: The covariance between 2 random variables X and Y is given by:

$$
\begin{aligned}
\mathrm{Cov}(X,Y) &= E\left[(X-E(X))(Y-E(Y))\right] \\
&= E\left[XY - E(X)Y - E(Y)X + E(X)E(Y)\right] \\
&= E(XY) - E(X)E(Y) - E(X)E(Y) + E(X)E(Y) \\
&= E(XY) - E(X)E(Y) \qquad (4.1.5)
\end{aligned}
$$

The variance of the sum $X+Y$ is given by:

$$
\begin{aligned}
\mathrm{Var}(X+Y) &= E\left[(X+Y-E(X+Y))^2\right] \\
&= E\left[(X+Y)^2 - 2(X+Y)E(X+Y) + E^2(X+Y)\right] \\
&= E[X^2 + 2XY + Y^2 - 2(XE(X) + YE(Y) + XE(Y) + YE(X)) \\
&\quad + E^2(X) + 2E(X)E(Y) + E^2(Y)] \\
&= E[(X^2 - 2XE(X) + E^2(X)) + (Y^2 - 2YE(Y) + E^2(Y)) \\
&\quad + 2(XY - XE(Y) - YE(X) + E(X)E(Y))] \\
&= E\left[(X-E(X))^2\right] + E\left[(Y-E(Y))^2\right] + 2E\left[(X-E(X))(Y-E(Y))\right] \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y) \qquad (4.1.6)
\end{aligned}
$$

4.1.4

Let $\sigma_X = \sqrt{\mathrm{Var}(X)}, \sigma_Y = \sqrt{\mathrm{Var}(Y)}$. Show that the correlation coefficient $\rho_{XY} = \dfrac{\mathrm{Cov}(X,Y)}{\sigma_X\sigma_Y}$ must stay between -1 and 1.

Solution: Consider the random variables centred by their means:

$$U = X - E(X) \qquad (4.1.7)$$
$$V = Y - E(Y) \qquad (4.1.8)$$

Since the expectation of a square is always non-negative, consider the following expression for any real number λ:

$$0 \leqslant E\left[(U+\lambda V)^2\right] = E(U^2) + 2\lambda E(UV) + \lambda^2 E(V^2) \qquad (4.1.9)$$

Consider this as a quadratic inequality in λ and it must hold for any real number λ. Therefore, its discriminant must be less than or equal to zero:

$$\begin{aligned}\Delta &= [2E(UV)]^2 - 4E(V^2)E(U^2) \leqslant 0 \\ \implies E^2(UV) &\leqslant E(U^2)E(V^2) \qquad (4.1.10)\end{aligned}$$

Substitute back the original terms:

$$\begin{aligned} E^2(UV) &\leqslant E(U^2)E(V^2) \\ \iff E^2\left[(X - E(X))(Y - E(Y))\right] &\leqslant E\left[(X - E(X))^2\right] E\left[(Y - E(Y))^2\right] \\ \iff \mathrm{Cov}^2(X, Y) &\leqslant \mathrm{Var}(X)\mathrm{Var}(Y) \\ \implies |\mathrm{Cov}(X, Y)| &\leqslant \sigma_X \sigma_Y \\ \implies -1 \leqslant \rho_{XY} &= \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y} \leqslant 1 \qquad (4.1.11)\end{aligned}$$

4.2

Information from 24 galaxies are shown in Table 1.

Table 1: Distance-velocity relation of galaxies in the original paper by Edwin Hubble

Nebulae	**Distance D (Mpc)**	**Radial velocity v_r (km/s)**
S. Mag	0.032	170
L. Mag	0.034	290
NGC 6822	0.214	-130
NGC 598	0.263	-70
NGC 221	0.275	-185
NGC 224	0.275	-220
NGC 5457	0.45	200
NGC 4736	0.5	290
NGC 5194	0.5	270
NGC 4449	0.63	200
NGC 4214	0.8	300
NGC 3031	0.9	-30
NGC 3627	0.9	650
NGC 4826	0.9	150
NGC 5236	0.9	500
NGC 1068	1.0	920
NGC 5055	1.1	450
NGC 7331	1.1	500
NGC 4258	1.4	500
NGC 4151	1.7	960
NGC 4382	2.0	500
NGC 4472	2.0	850
NGC 4486	2.0	800
NGC 4649	2.0	1000

4.2.1

Fit a linear relation between D and $v_{\rm r}$ in the form of $v_{\rm r} = H_0 D$ to obtain the Hubble constant H_0. Visualize the data set, together with the fitting result, in a scattered plot of distance D vs. radial velocity $v_{\rm r}$ on a linear scale.

Solution: We perform a linear regression analysis. The optimal Hubble constant $\hat{H}_0$ minimizes the sum of squared residuals and can be derived analytically:

$$\begin{aligned} \hat{H}_0 &= \arg\min_{\hat{H}_0} \sum_{i=1}^{n} (v_{{\rm r},i} - \hat{H}_0 D_i)^2 \\ &= \frac{\sum_{i=1}^{n} D_i v_{{\rm r},i}}{\sum_{i=1}^{n} D_i^2} \\ &\approx 417.84 \text{ km/s/Mpc} \end{aligned} \tag{4.2.1}$$

The scatter plot of distance D vs. radial velocity $v_{\rm r}$ with the best-fit line is shown in Figure 1.

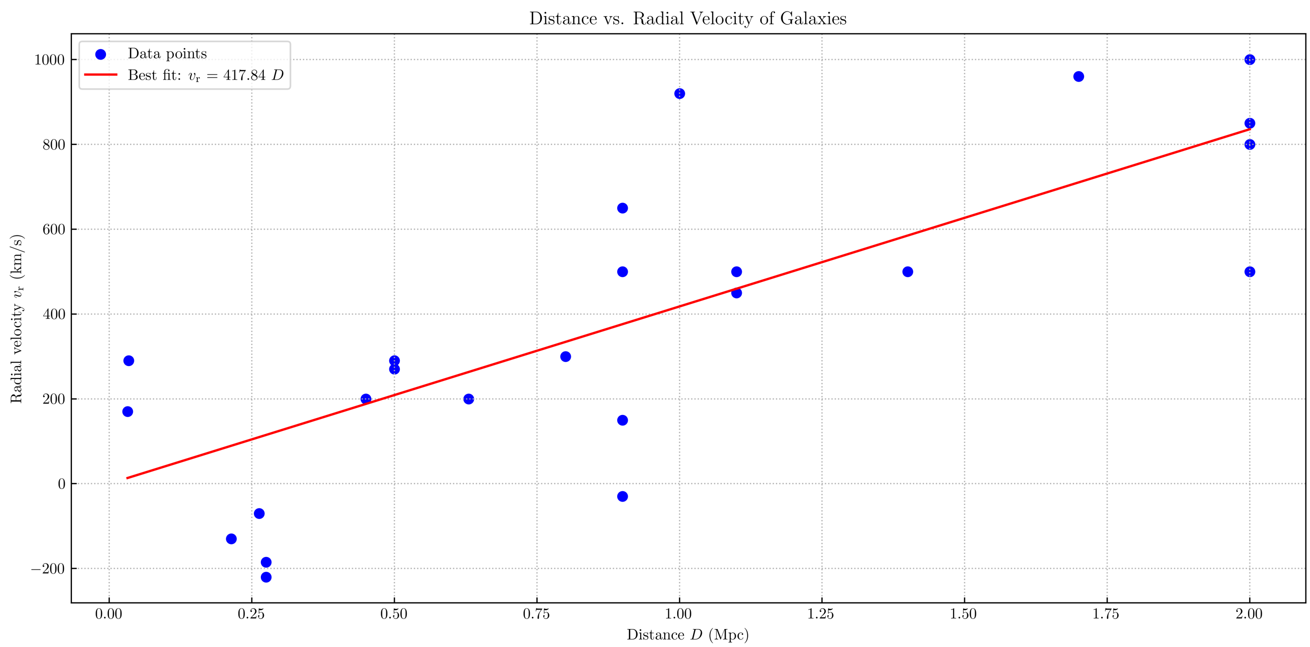

Figure 1: Scatter plot of distance D (Mpc) vs. radial velocity $v_{\rm r}$ (km/s) with the fitted linear relation $v_{\rm r} \approx 417.84\ D$.

4.2.2

Use bootstrap to estimate the 95% and 99% confidence intervals of the fitted Hubble constant.

Solution: We first generate a large number of bootstrap samples. Each bootstrap sample is created by randomly sampling with replacement from the original 24 galaxy data points. Then for each bootstrap sample, perform the same linear regression $v_{\rm r} = \hat{H}_0 D$ to obtain a bootstrap estimate $\hat{H}_0$. To construct the 95% confidence interval, determine the 2.5th and 97.5th percentiles of the bootstrap distribution of $\hat{H}_0$. Similarly, to construct the 99% confidence interval, determine the 0.5th and 99.5th percentiles of the bootstrap distribution of $\hat{H}_0$. A histogram is displayed in Figure 2 showing the distribution of the bootstrap estimates of the Hubble constant $\hat{H}_0$. The 95% confidence interval of the fitted Hubble constant is [338.46, 496.14] km/s/Mpc and the 99% confidence interval is [310.28, 524.06] km/s/Mpc.

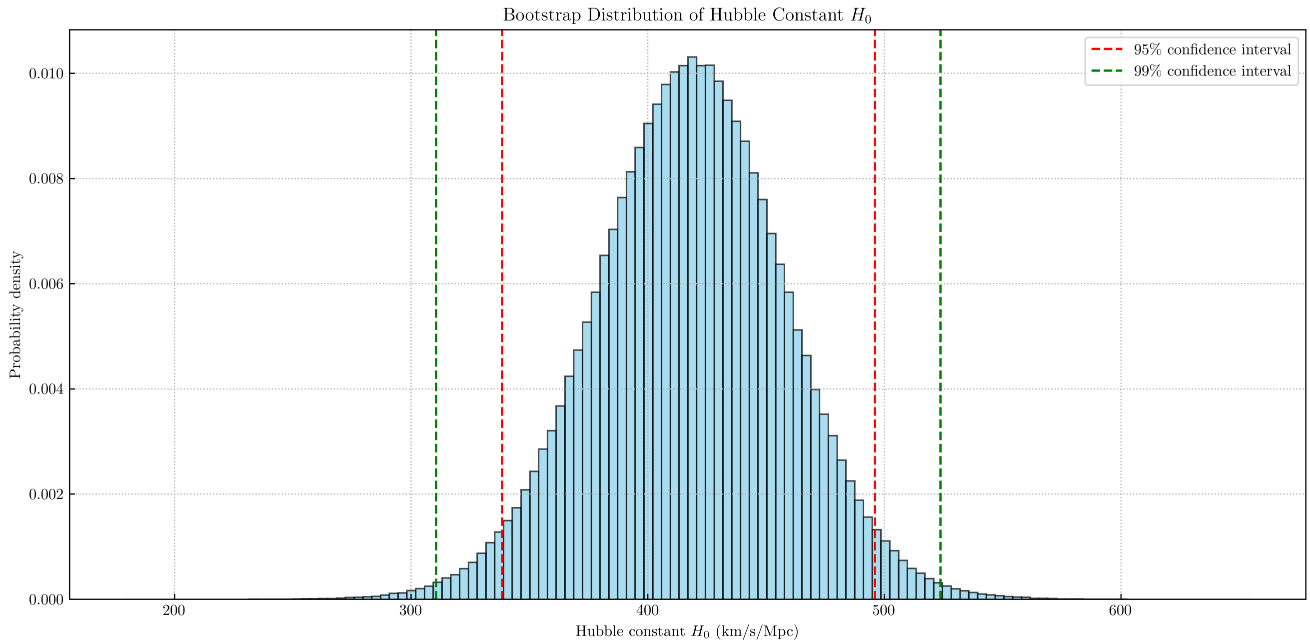

Figure 2: Bootstrap distribution of Hubble constant H_0.

4.2.3

Calculate sample Pearson correlation coefficient r and use bootstrap to determine its 95% and 99% confidence interval.

Solution: The formula for sample Pearson correlation coefficient r is:

$$r = \frac{\sum_{i=1}^{n}(D_i - \overline{D})(v_{\mathrm{r},i} - \overline{v}_r)}{\sqrt{\sum_{i=1}^{n}(D_i - \overline{D})^2}\sqrt{\sum_{i=1}^{n}(v_{\mathrm{r},i} - \overline{v}_r)^2}} \approx 0.78707 \tag{4.2.2}$$

where $\overline{D}$ and $\overline{v_r}$ are the sample means of D and v_{r}, respectively. We first generate a large number of bootstrap samples by resampling with replacement from the original dataset. Then for each bootstrap sample, calculate Pearson correlation coefficient $\hat{r}$. Next, identify different percentiles of the bootstrap distribution of $\hat{r}$. Figure 3 shows the distribution of the bootstrap estimates of the sample Pearson correlation coefficient r. The 95% and 99% confidence intervals are [0.62731, 0.90227] and [0.55744, 0.92681], respectively.

4.2.4

Use the F-test to assess whether the linear relation holds (assuming the residuals are Gaussian) for significance level 0.05 and 0.01, respectively, and give the p-value. (Hint: this is to compare the variance between a constant fit and a linear fit.)

Solution: We compare 2 models:

- Null hypothesis: The data follows a **constant model** $v_{\mathrm{r}} = \mu$, where μ is the mean radial velocity.

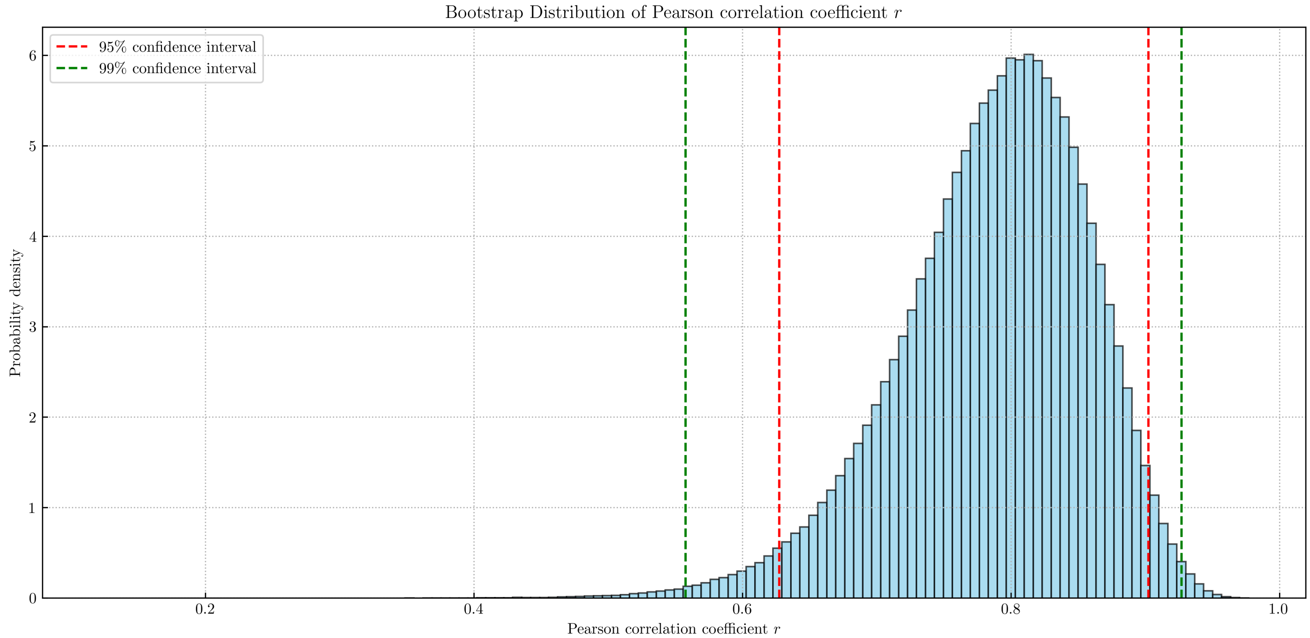

Figure 3: Bootstrap distribution of sample Pearson correlation coefficient r.

- Alternative hypothesis: The data follows a **linear model** $v_{\mathrm{r}} = H_0 D$.

The F-statistic is calculated by comparing the variances of the 2 models. Specifically, it assesses whether the reduction in the residual sum of squares from the constant model to the linear model is significant.

$$F = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(v_{\mathrm{r},i} - \mu)^2}{\frac{1}{n-1}\sum_{i=1}^{n}(v_{\mathrm{r},i} - \hat{H}_0 D_i)^2} = \frac{\sum_{i=1}^{n}(v_{\mathrm{r},i} - \mu)^2}{\sum_{i=1}^{n}(v_{\mathrm{r},i} - \hat{H}_0 D_i)^2} \approx 2.60622 \tag{4.2.3}$$

where $\hat{H}_0$ is calculated using Equation 4.2.1. Using the calculated F-statistic and the degrees of freedom ($\mathrm{df1} = \mathrm{df2} = 24 - 1 = 23$), determine the p-value from the F-distribution table:

$$p = P(F > 2.60622) \approx 0.013 \tag{4.2.4}$$

Since $p \approx 0.013 < 0.05$, reject the null hypothesis at significance level $\alpha = 0.05$. However, since $p \approx 0.013 > 0.01$, there is no strong evidence to support the existence of a linear relationship between D and v_{r} at significance level $\alpha = 0.01$.