Statistics & Numerical Method, Problem Set 4 (due Nov. 26, 2024)

1. Check your understanding (8 pts)

Suppose $X_1, \cdots, X_n$ are independent identically distributed random variables [IIDs, for questions (1), (2)]. Let $\overline{X}$ be their sample average, and use $E(X)$ and $\mathrm{Var}(X)$ to denote the expectation and variance of these variables (they both exist). Further let X and Y are two random variables [for questions (3), (4)], and the expectation and variance both exist for these variables.

(1). Prove that $\mathrm{Var}(\overline{X})=\mathrm{Var}(X)/n$. (2 pts)

(2). Define $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Show that $E(S^2)=\mathrm{Var}(X)$ thus S^2 is an unbiased estimate of $\mathrm{Var}(X)$. (2 pts)

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

(4). Let $\sigma_X = \sqrt{\mathrm{Var}(X)}, \sigma_Y = \sqrt{\mathrm{Var}(Y)}$. Show that the population correlation coefficient $\rho_{XY} = \mathrm{Cov}(X,Y)/\sigma_X\sigma_Y$ must stay between -1 and 1. (2 pts)

2. The Hubble Law (16 pts)

The well-known Hubble's law describes the observation that galaxies are moving away from the Earth at speeds proportional to their distance, which is a cornerstone of physical cosmology. The proportionality, H_0, is known as the Hubble constant. It's original discovery, widely attributed to Edwin Hubble (1929, PNAS, 15, 168), who employed very rough measurements of galaxy distance and redshifts (converted to radial velocity) known at the time, and fitted it into a linear relation. The data are shown in Table 1, with information from 24 galaxies. We are going to employ a number of statistical tools to reassess his discovery. Please note that the fitting result based on this data is larger than the modern value of the Hubble constant by nearly one order of magnitude. Nevertheless, it does not prevent us from practicing some statistics, and also bare in mind that you don't often have the luxury to play with beautiful statistics in real research data!

(1). Fit a linear relation between D and v_r in the form of $v_r = HD$ to obtain the Hubble constant H. (2 pts) Visualize the data set, together with the fitting result, in a scattered plot of distance D vs. radial velocity v_r on linear scale. (1 pt)

(2). Use bootstrap to estimate the 95% and 99% confidence intervals of the fitted Hubble constant. (5 pts)

(3). Calculate Pearson's sample correlation coefficient r, and use bootstrap to determine its 95% and 99% confidence interval. (3 pts)

(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.) You may find standard tables for the F-test online, or simply numerically compute the CDF of the F-distribution on your own. (7 pts)

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

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