Problem 1

$X_1, X_2, \ldots$ are i.i.d. random variables, each with density $f(x)$. X_i measures the annual number of precipitations at a specific location. The distribution of the number of years that pass until the rains of the first year, X_1, are surpassed for the first time will be computed in this problem.

This situation will be described by a geometric distribution. If the probability of success on each trial is p, then the probability that the k^{th} trial is the first success is given by

$$P(X = k) = (1-p)^{k-1}p. \tag{1}$$

Next, it is apparent that

$$p = \int_{x_1}^{\infty} f(x)\mathrm{d}x = P(x_1 < X_i), \quad i \in [2, \infty], \tag{2}$$

the probability that X_i is larger than x_1. The probability is the same for all $i \in [2, \infty]$ since the density of each X_i is $f(x)$. Hence, the distribution of the number of years until the rains of the first year are surpassed for the first time is given by equation (1), where $k = 1, 2, \ldots$ are the years that pass and the probability p that each year will surpass the rain in X_1 is given by equation (2).

Problem 2

$\overline{X}_1$ and $\overline{X}_2$ are sample means calculated from two independent samples of size n from a population with variance σ^2. In this problem I will find the smallest value of n that guarantees that

$$P\left(|\overline{X}_1 - \overline{X}_2| < \frac{\sigma}{5}\right) \geq 0.99.$$

First of all, Chebyshev's inequality states that

$$P(|X| \geq a) \leq \frac{\mathrm{E}(X^2)}{a^2}, \tag{3}$$

if X is a random variable and $\mathrm{E}(X^2) < \infty$. This equation can be restated as

$$P(|X| < a) \geq 1 - \frac{\mathrm{E}(X^2)}{a^2}. \tag{4}$$

Letting $X := \overline{X}_1 - \overline{X}_2$ and $a := \sigma/5$ gives

$$P(|\overline{X}_1 - \overline{X}_2| < \frac{\sigma}{5}) \geq 1 - \frac{\mathrm{E}((\overline{X}_1 - \overline{X}_2)^2)}{(\frac{\sigma}{5})^2}.$$

$\mathrm{E}(X^2)$ can be calculated with the moment generating function

$$M_X(t) = \exp\left(\mu t + \frac{\sigma_*^2 t^2}{2}\right) = \exp\left(\frac{\sigma^2 t^2}{n}\right),$$

since $X \overset{n\to\infty}{\sim} N(\mu = 0, \sigma_*^2 = \frac{2\sigma^2}{n})$ (from the Central Limit Theorem) and $\mathrm{E}(X^2) = M_X''(0)$.
Thus,

$$\begin{aligned}
M_X'(t) &= \frac{2t\sigma^2}{n}\exp\left(\frac{\sigma^2 t^2}{n}\right) \\
M_X''(t) &= \frac{2\sigma^2}{n}\exp\left(\frac{\sigma^2 t^2}{n}\right) + \left(\frac{2t\sigma^2}{n}\right)^2 \exp\left(\frac{\sigma^2 t^2}{n}\right) \\
\mathrm{E}(X^2) &= M_X''(0) = \frac{2\sigma^2}{n}.
\end{aligned}$$

Hence,

$$P(|\overline{X}_1 - \overline{X}_2| < \frac{\sigma}{5}) \geq 1 - \frac{2\sigma^2/n}{(\frac{\sigma}{5})^2} = 1 - \frac{50}{n} \overset{!}{\geq} 0.99,$$

which gives $50/n \leq 0.01 \implies n \geq \frac{50}{0.01} = 5000$. Thus, the smallest value of n to obtain the result is $n = 5000$.

Comment on the result: "Only" 5000 samples are needed from a population to insure that the difference between two independent sample means is $\sigma/5$ with a probability higher than 0.99. I think this was surprisingly few samples needed for such a "strict" result.

Problem 3

$U_i,\ \ i = 1, 2, \ldots$ are independent random variables with distribution $U(0, 1)$. X is a random variable with distribution

$$P(X = x) = \frac{1}{(e-1)x!},\ \ x = 1, 2, 3, \ldots. \tag{5}$$

In this problem I will give the distribution of $Z = \min\{U_1, \ldots, U_X\}$. A hint states that $Z|X = x$ is the first order statistic from a sample of size x from $U(0, 1)$. Following the hint, we know that $U_i,\ \ i = 1, 2, \ldots$ has to be greater than a given scalar t iff $Z|X = x$ is greater than t. Thus,

$$\begin{aligned}
F_{Z|X=x}(t) &= P(Z|X = x \leq t) = 1 - P(Z|X = x > t) \\
&= 1 - P(\min\{U_1, \ldots, U_x\} > t) = 1 - P(U_1 > t, \ldots, U_x > t) \\
&\overset{Indep.}{=} 1 - P(U_1 > t) \cdot \ldots \cdot P(U_x > t) = 1 - (1 - P(U_1 \leq t)) \cdot \ldots \cdot (1 - P(U_x \leq t)) \\
&= 1 - (1 - F_{U_1})^x.
\end{aligned}$$

Hence, we know the distribution of Z, *given* the sample size x. Now, $F_Z(t)$ needs to be calculated from this[1].

Problem 4

$X_1, X_2, \dots X_n$ is a simple random sample from a population with density $f_X(x)$, $-\infty < x < \infty$. For $i = 1, 2, \dots, n$, let $U_i = F_X(X_i)$, being $F_X(x) = \int_{-\infty}^{x} f_x(t)\mathrm{d}t$. In this problem I will find the distribution of variable U_i. The transformation $U_i = F_X(X_i)$ is known as the *probability integral transformation*.

The CDF of U_i can be calculated as

$$F_U(u) = P(U \leq u) = P(F_X(X) \leq u) = P(X \leq F_X^{-1}(u)) = F_X(F_X(u)^{-1}) = u,$$

assuming that F_X is a bijection[2]. Thus, $U_i \sim U(0, 1)$.

Problem 5

Prove that the following are exponential families and describe the natural parametric space for each one of them[3].

Recall the definition of the exponential family

$$f(x|\theta) = h(x)c(\theta)\exp\left(\sum_{j=1}^{k} w_j(\theta)t_j(x)\right) \tag{6}$$

and the natural parametric space

$$\Theta = \left\{\eta \in \mathbb{R}^k : \int_{-\infty}^{\infty} h(x)\exp\left(\sum_{j=1}^{k}\eta_j t_j(x)\right)\mathrm{d}x < \infty\right\}. \tag{7}$$

(a) Normal family with one of the parameters μ or σ known.

The normal family has the form

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Now, when μ is known, σ is the only parameter. In this case, we can define the functions

[1] Stuck here.
[2] I do not think this is an assumption that can be made in general. Can it be made in this case?
[3] (b), (c) and (e) are not included here.

$$h(x) := 1,\ c(\theta) := \frac{1}{\sigma\sqrt{2\pi}}$$
$$w_1(\theta) := -\frac{1}{2\sigma^2},\ t_1(x) := (x-\mu)^2,$$

for $k = 1$. With these definitions it is trivial to see that the normal family is an exponential family. The natural parametric space can be found by

$$\int_{-\infty}^{\infty} h(x) \exp\left(\sum_{j=1}^{k} \eta_j t_j(x)\right) \mathrm{d}x = \int_{-\infty}^{\infty} \exp\left(\eta_1 (x-\mu)^2\right) \mathrm{d}x < \infty,$$

which holds when $\eta_1 < 0$. Thus the natural parametric space is $\Theta = \{\eta \in \mathbb{R} : \eta < 0\} = \mathbb{R}^-$.

Next, when σ is known, μ is the only parameter. In this case, we can define the functions

$$h(x) := \frac{1}{\sigma\sqrt{2\pi}},\ c(\theta) := 1$$
$$w_1(\theta) := -\frac{\mu^2}{2\sigma^2},\ t_1(x) := -1$$
$$w_2(\theta) := -\frac{\mu}{2\sigma^2},\ t_2(x) := -2x$$
$$w_3(\theta) := \frac{1}{2\sigma^2},\ t_3(x) := -x^2,$$

for $k = 3$. These functions show that the normal family is an exponential family also when μ is the only parameter. The calculations of the natural parametric space is tedious and will be skipped here.

(b) Gamma family with one of the parameters α or β known.

(c) Beta family with one of the parameters α or β known.

(d) Poisson family.

The Poisson family has the form $f(x;\lambda) = \frac{\lambda^x e^{-\lambda}}{x!} = \exp(-\lambda)\exp(x\log\lambda - \log x!)$. Define the functions

$$h(x) := 1,\ c(\theta) := \exp(-\lambda)$$
$$w_1(\theta) := \log\lambda,\ t_1(x) := x$$
$$w_2(\theta) := 1,\ t_2(x) := -\log x!,$$

for $k = 2$. These functions show that the Poisson family is an exponential family, since the distribution function can be written in the required form. Calculations of the natural parametric space lead to the fact that

$$\int_{-\infty}^{\infty} e^{\eta_1 x}(x!)^{-\eta_2}\mathrm{d}x < \infty,$$

which I believe holds when $\eta_2 > 0, \forall \eta_1$.

(e) Negative Binomial with the parameter r known and $0 < p < 1$.

Problem 6

$X_1, \ldots, X_n$ are independent random variables from the exponential family of distributions, expressed in terms of the natural parametric space. Prove that the joint distribution of the n variables also belongs to the exponential family.

Proof. First of all, the definition of the exponential family in terms of the natural parametric space is

$$f(x|\eta) = h(x)c^*(\eta)\exp\left(\sum_{j=1}^{p}\eta_j t_j(x)\right). \tag{8}$$

Independent random variables implies that the joint distribution can be written as

$$\begin{aligned}
f(\mathbf{X}|\eta) &= \prod_{i=1}^{n} f_{X_i}(x_i|\eta) \\
&= \prod_{i=1}^{n} h(x_i)c^*(\eta)\exp\left(\sum_{j=1}^{p}\eta_j t_j(x_i)\right) \\
&= \prod_{i=1}^{n} h(x_i)\cdot(c^*(\eta))^n \cdot \exp\left(\sum_{j=1}^{p}\eta_j \sum_{i=1}^{n} t_j(x_i)\right) \\
&= g(x)\cdot z^*(\eta)\cdot\exp\left(\sum_{j=1}^{p}\eta_j p_j(x)\right),
\end{aligned}$$

where I have defined $g(x) := \prod_{i=1}^{n} h(x_i)$, $z^*(\eta) := (c^*(\eta))^n$ and $p_j(x) := \sum_{i=1}^{n} t_j(x_i)$. □

Problem 7

$X_1, \ldots, X_n$ are independent random variables such that $X_1 \sim \mathrm{Exp}(1/(i\theta)), \mathrm{E}(X_i) = i\theta, \theta > 0$. Prove that the family of the joint distributions of the n variables is an exponential family.

Proof. Independent random variables implies that the joint distribution can be written as

$$\begin{aligned} f(\mathbf{X}) &= \prod_{i=1}^{n} f_{X_i}(x_i|\theta) \\ &= \prod_{i=1}^{n} \frac{1}{i\theta} \exp\left(-\frac{x}{i\theta}\right) \\ &= \prod_{i=1}^{n} \frac{1}{i\theta} \cdot \exp\left(-x \sum_{i=1}^{n} \frac{1}{i\theta}\right) \\ &= h(x)c(\theta) \exp\left(t(x) \sum_{i=1}^{n} w_i(\theta)\right), \end{aligned}$$

where I have defined $h(x) := 1$, $c(\theta) := \prod_{i=1}^{n} \frac{1}{i\theta}$, $t(x) := -x$ and $w_j(\theta) := \frac{1}{i\theta}$, $j = 1, \ldots, n$. □

Problem 8

In this problem I will prove that if $f(x)$ is a symmetric density function centered at 0, then the median M of the density

$$\frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right)$$

equals μ.

Proof. M is the value such that

$$\int_{-\infty}^{M} f(x)\mathrm{d}x = \frac{1}{2}.$$

In this case, this gives

$$\int_{-\infty}^{M} \frac{1}{\sigma} f\left(\frac{x-\mu}{\sigma}\right) \mathrm{d}x = \frac{1}{2}.$$

The change of variables $y = \frac{x-\mu}{\sigma}$, $\mathrm{d}y = \frac{1}{\sigma}\mathrm{d}x$ gives

$$\int_{-\infty}^{(M-\mu)/\sigma} f(y)dy = \frac{1}{2}.$$

Since f is symmetric around 0, we know that

$$\int_{-\infty}^{0} f(y)\mathrm{d}y = \frac{1}{2},$$

which implies that $\frac{M-\mu}{\sigma} = 0 \implies M = \mu$.

□

Problem 9

$X_1, X_2, \ldots, X_n$ are i.i.d. $N(\mu_1, \sigma^2)$. $Y_1, \ldots, Y_n$ are i.i.d. $N(\mu_2, 3\sigma^2)$. Moreover, $\mu_1, \mu_2 \in \mathbb{R}, \sigma \in \mathbb{R}^+$ and X, Y independent. Denote by $\overline{X}, \overline{Y}, S_{1n}^2, S_{2n}^2$ the sample means and sample variances of X and Y. Define $T_n = S_{1n}^2 + \frac{1}{3}S_{2n}^2$ for $n \geq 2$.

(a) Let $V_n := \frac{1}{2\sigma}\sqrt{n}(\overline{X} - \overline{Y} - \mu_1 - \mu_2)$. The distribution of this random variable is obviously normal, since $\overline{X} \sim N(\mu_1, \sigma^2/n)$ and $\overline{Y} \sim N(\mu_2, 3\sigma^2/n)$, under the assumptions. The expectation and variance are

$$\begin{aligned}
\mathrm{E}[V_n] &= \frac{\sqrt{n}}{2\sigma}(\mu_1 - \mu_2 - \mu_1 + \mu_2) = 0 \\
\mathrm{Var}[V_n] &\overset{Indep.}{=} \frac{n}{4\sigma^2}(\mathrm{Var}[\overline{X}] + \mathrm{Var}[\overline{Y}]) = \frac{n}{4\sigma^2} \cdot \frac{4\sigma^2}{n} = 1.
\end{aligned}$$

Hence, $V_n \sim N(0, 1)$.

(b) Let $T_n := S_{1n}^2 + \frac{1}{3}S_{2n}^2$, for $n \geq 2$. Here I will show that $(n-1)T_n/\sigma^2 \sim \chi^2_{2(n-1)}$. First of all, recall that $(n-1)S_{1n}^2/\sigma^2 \sim \chi^2_{n-1}$ and $(n-1)S_{2n}^2/(3\sigma^2) \sim \chi^2_{n-1}$. Thus

$$\frac{(n-1)}{\sigma^2}T_n = \frac{(n-1)}{\sigma^2}\left(S_{1n}^2 + \frac{1}{3}S_{2n}^2\right) = \frac{(n-1)}{\sigma^2}S_{1n}^2 + \frac{(n-1)}{3\sigma^2}S_{2n}^2 \sim \chi^2_{2(n-1)},$$

where the last equation holds since the sum of two independent χ_k^2 distributed variables has the distribution χ_{2k}^2. I will show this with the moment generating function of χ_k^2. Let $A_1, A_2 \sim \chi_k^2$ and A_1, A_2 independent. Then

$$M_{(A_1+A_2)}(t) \overset{Indep.}{=} M_{A_1}(t) \cdot M_{A_2}(t) = (1-2t)^{-k/2 \cdot 2} = (1-2t)^{-k} = M_B(t), \tag{9}$$

where $A_1 + A_2 = B \sim \chi_{2k}^2$.

(c) Are V_n and T_n independent? The answer to this question is *yes*, they are independent. This can be seen by looking at the definitions of the two random variables, in conjunction with the following result (Fischer's theorem)

Result 1. *Let* $X_1, \dots, X_n \overset{i.i.d}{\sim} N(\mu, \sigma^2)$. *Then, the statistics*

$$\overline{X}$$

and

$$(X_1 - \overline{X}, \dots, X_n - \overline{X})$$

are independent.

A proof can be constructed based on moment generating functions[4].

Recall the definitions of the statistics

$$V_n = \frac{1}{2\sigma}\sqrt{n}(\overline{X} - \overline{Y} - \mu_1 - \mu_2),$$
$$T_n = S_{1n}^2 + \frac{1}{3}S_{2n}^2.$$

The above result implies that $\overline{X}, S_{1n}^2$ are independent. Similarly, it implies that $\overline{Y}, S_{2n}^2$. Moreover, both these pairs of statistics must be independent, since X, Y are independent. Thus, since V_n and T_n are linear combinations of these statistics, they must be independent as well.

(d) In this problem I will show that $U_n := \sqrt{n}(\overline{X} - \overline{Y} - \mu_1 + \mu_2)/\sqrt{2T_n} \sim t_{2(n-1)}$. Recall that the definition of the Student's t-distribution states that $Z/\sqrt{V/\nu} \sim t_\nu$, when $Z \sim N(0,1)$, $V \sim \chi_\nu^2$ and Z, V independent. The result follows directly from this definition, using the above results

$$t_{2(n-1)} \sim T = \frac{V_n}{\sqrt{\frac{(n-1)}{\sigma^2}T_n/2(n-1)}} = \frac{\frac{1}{2\sigma}\sqrt{n}(\overline{X} - \overline{Y} - \mu_1 - \mu_2)}{\left(\frac{T_n}{2\sigma^2}\right)^{1/2}} = \frac{\sqrt{n}(\overline{X} - \overline{Y} - \mu_1 - \mu_2)}{\sqrt{2T_n}}.$$

(e) In this problem I will show that $T_n \xrightarrow{P} 2\sigma^2$, i.e. that T_n converges in probability to $2\sigma^2$.

Proof. The definition states that T_n converges in probability to $2\sigma^2$ if, for every $\epsilon > 0$,

$$\lim_{n\to\infty} P(|T_n - 2\sigma^2| \geq \epsilon). \tag{10}$$

Using Chebyshev's inequality yields that

$$(|T_n - 2\sigma^2| \geq \epsilon) \leq \frac{\mathrm{E}[(T_n - 2\sigma^2)]}{\epsilon^2} = \frac{1}{\epsilon^2}(\mathrm{Var}_{2\sigma^2}[T_n] + (\mathrm{B}_{2\sigma^2}(T_n))^2), \tag{11}$$

[4]Can be found in Gómez (UPC) and Sánchez (UB) IEA Unit 0 in Advanced Statistical Inference at UPC.

where the last equality holds because of the relationship between MSE, Bias and Variance. The bias of T_n is zero[5]. Moreover, the variance of T_n is

$$\begin{aligned}
\text{Var}\left(S_{1n}^2 + \frac{1}{3}S_{2n}^2\right) &\overset{Indep.}{=} \text{Var}(S_{1n}^2) + \text{Var}\left(\frac{1}{3}S_{2n}^2\right) \\
&= \frac{2\sigma^4}{n-1} + \frac{1}{9}\text{Var}\left(\frac{(n-1)}{3\sigma^2}S_{2n}^2\frac{3\sigma^2}{(n-1)}\right) \\
&= \frac{2\sigma^4}{n-1} + \frac{1}{9}\frac{9\sigma^4}{(n-1)^2}\cdot 2(n-1) \\
&= \frac{4\sigma^4}{n-1}.
\end{aligned}$$

This means that

$$\lim_{n\to\infty} P(|T_n - 2\sigma^2| \geq \epsilon) \leq \lim_{n\to\infty} \frac{4\sigma^2}{n-1} = 0, \;\; \forall \epsilon > 0,$$

which implies that $\lim_{n\to\infty} P(|T_n - 2\sigma^2| \geq \epsilon) = 0, \;\; \forall \epsilon > 0$. Hence, we have shown that T_n converges in probability to $2\sigma^2$. □

(f) In this problem I will show that $U_n \xrightarrow{\mathcal{L}} N(0,1)$, i.e. that U_n converges in law to $N(0,1)$. As far as I can understand, convergence in law is also called convergence in distribution. The definition of convergence in distribution states that a sequence of random variables $U_i, \;\; i \in \mathbb{N}$ converges in distribution to a random variable X if

$$\lim_{n\to\infty} F_{U_n}(x) = F_X(x), \tag{12}$$

at all points x where $F_X(x)$ is continuous. This definition is not really an appropriate tool to use to prove convergence in law[6].

Problem 10

In this problem, the properties that were shown in the previous problem are illustrated using R[7].

[5] Skipped showing this here.

[6] I am stuck here however.

[7] The simulation of problem (b) does not give the expected results, despite using the same code as in (d). Did not manage to figure out why when writing this. Also, how can I check if two statistics are independent via R? Is it possible to simulate (e) (convergence in probability) and (f) (convergence in law) also?

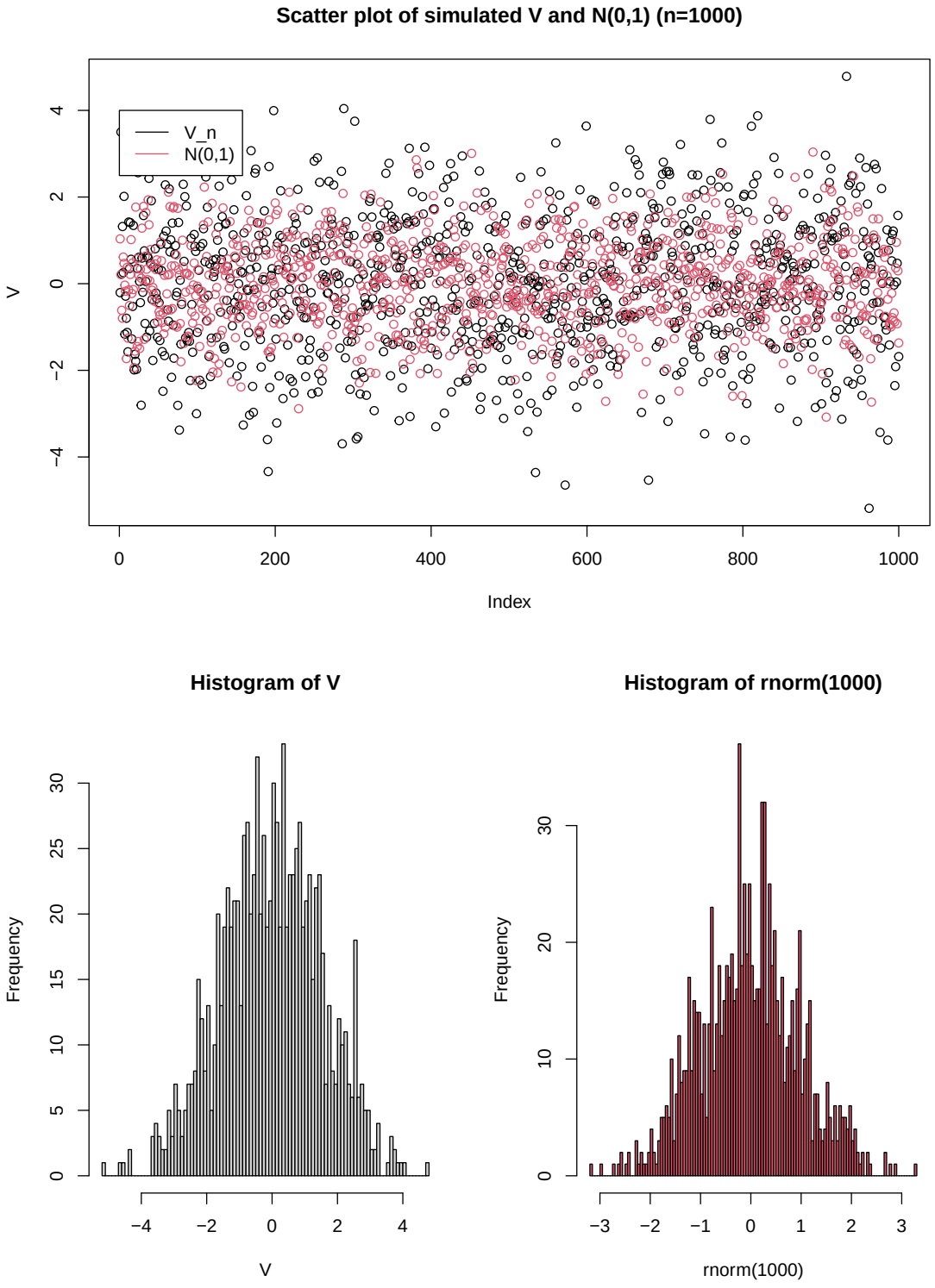

Figure 1: (a) Looks like the simulated samples of V_n are normally distributed.

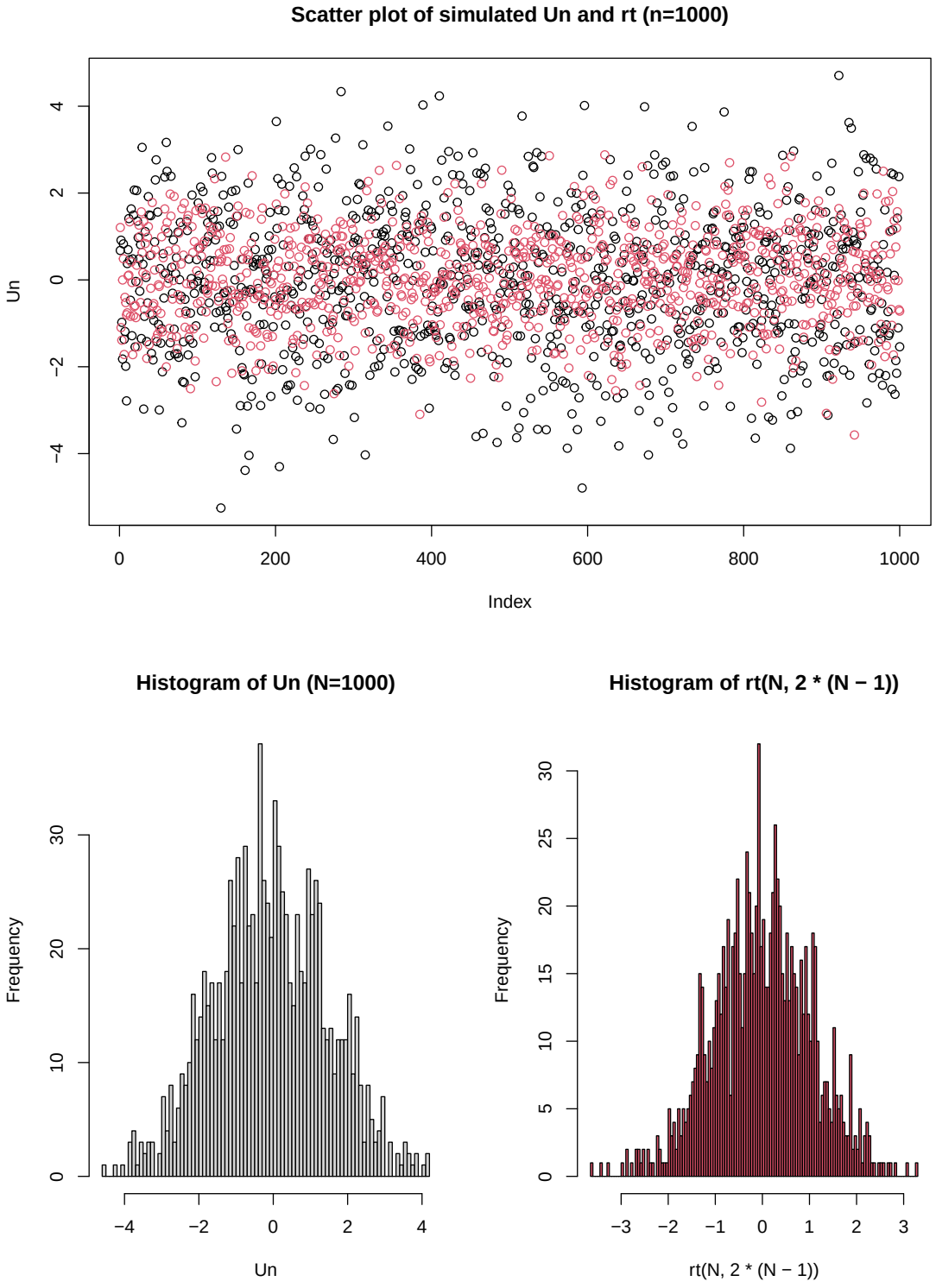

Figure 2: (d) Looks like the simulated samples of U_n are distributed as Student's $t_{2(n-1)}$.