EPS 659a — Time Series in Geoscience

Problem Set Two

due Friday, Sept 17, 2021

Problem 1: R Functions for Normal (Gaussian) Distribution

The normal (Gaussian) probability density function with mean μ, standard deviation σ and variance σ^2 can be written as

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right)$$

is represented by four R functions. The function `rnorm` returns a vector of random-variable realizations of $p(X)$, as we saw in Problem Set 1. The function `dnorm` returns a vector of $p(X)$ values, corresponding to an input vector of random-variable values X_k. The function `pnorm` returns a vector of values for the cumulative distribution $P(X)$, corresponding to an input vector of random-variable values X_k. The function `qnorm` returns a vector of values X_k for an input vector of values for the cumulative distribution $P(X)$. On Canvas you can find `probset2_template.Rmd` that will show how these functions operate within RStudio.

a) For $\mu = 1$ and $\sigma = 2$, plot the Gaussian PDF and CDF for $-6 < X < 6$.

b) For $\mu = 1$ and $\sigma = 2$, confirm the normalization condition $\int_{-\infty}^{\infty} p(X)dX = 1$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$.

c) For $\mu = 1$ and $\sigma = 2$, confirm the mean of the PDF $\int_{-\infty}^{\infty} Xp(X)dX = \mu$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$.

d) For $\mu = 1$ and $\sigma = 2$, confirm the variance of the PDF $\int_{-\infty}^{\infty} (X-\mu)^2 p(X)dX = \sigma^2$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$.

e) For $\mu = 0$ and $\sigma = 1$, the sample markdown file `probset2_template.Rmd` contains the following commands:

```
x <- seq(-5.0,5.0, by=0.01)
dd <- dnorm(x,sd=1)
nn <- 10000
xx <- rnorm(nn,sd=1)
hist(xx,breaks=seq(-6.0,6.0,by=0.2))
ddnn <- nn*dd/5.0
lines(x,ddnn)
```

Explain why the sixth line of R code is useful: `ddnn <- nn*dd/5.0`.

f) For the Gaussian with $\mu = 1$ and $\sigma = 2$, what are the bounding values of the 10% and 90% percentiles of the distribution?

Problem 2: R Functions for χ^2 Distribution

The χ^2 probability density function with ν degrees-of-freedom can be written as

$$p(X) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} X^{(\nu/2)-1} e^{-X/2}$$

where $\Gamma(K)$ is the gamma function.

When K is a positive integer, $\Gamma(K) = (K-1)! = (K-1)(K-2)(K-3)\ldots 3 \cdot 2 \cdot 1$.

When K is a positive half-integer, e.g., $K = 5/2$, $\Gamma(K) = (K-1)(K-2)(K-3)\ldots(3/2) \cdot (1/2) \cdot \sqrt{\pi}$.

The random variable X is a sum of ν squared Gaussian random variables. (This interpretation applies only if the parameter ν is a positive integer. In practice this PDF is well-defined for real-valued $\nu > 0$, not only the integers.) The interpretation for integers indicates that the χ^2 PDF is appropriate for the statistics of the variance represented by a finite collection of Gaussian realizations. The formula above for the PDF is appropriate for the zero-mean, unit-variance Gaussian PDF as the source of the squared random variables. The random variable X only can take non-negative values. The χ^2 distribution is key to understanding the statistics of least-squares data-fitting, and also the statistics of spectrum estimates in time-series methods.

The χ^2 distribution is represented by four R functions. The function `rchisq` returns a vector of random-variable realizations of $p(X)$. The function `dchisq` returns a vector of $p(X)$ values, corresponding to an input vector of random-variable values X_k. The function `pchisq` returns a vector of values for the cumulative distribution $P(X)$, corresponding to an input vector of random-variable values X_k. The function `qchisq` returns a vector of values X_k for an input vector of values for the cumulative distribution $P(X)$.

a) For $\nu = 2$, show that the χ^2 PDF reduces to a normalized exponential function, that is, normalized so that $\int_0^\infty p(X)dX = 1$.

b) For $\nu = 4$, plot the PDF for $X \in [0, 10]$, compute the peak values of the PDF using calculus and mark it on the graph.

c) For $\nu = 4$, confirm the normalization condition $\int_0^\infty p(X)dX = 1$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$. Compute the expected value $\langle X \rangle$ numerically as well, to confirm its theoretical value $\langle X \rangle = \nu$.

d) For $\nu = 10$, plot the PDF for $X \in [0, 30]$, compute the peak values of the PDF using calculus and mark it on the graph.

e) For ν = 2,4,6,8, and 10, use the CDF function `pchisq` to compute the probability of $X > 2\nu$ for each value of ν. Do these relative deviations from the expected value $\langle X \rangle = \nu$ become less probable or more probable as ν increases?

Problem 3: R Functions for Log-Normal Distribution

The log-normal probability density function is a PDF in which the logarithm of the random variable X follows Gaussian probability, with mean μ, standard deviation σ and variance σ^2. This PDF applies for $X > 0$ as

$$p(X) = \frac{1}{\sqrt{2\pi}\sigma X} \exp\left(-\frac{(\ln X - \mu)^2}{2\sigma^2}\right)$$

where $\ln X$ is the natural log of X, that is, $\ln 1 = 0$, $\ln e = 1$, and $\ln X \to -\infty$ as $X \to 0+$. The log-normal PDF is represented by four R functions. The function `rlnorm` returns a vector of

random-variable realizations of $p(X)$, as we saw in Problem Set 1. The function `dlnorm` returns a vector of $p(X)$ values, corresponding to an input vector of random-variable values X_k. The function `plnorm` returns a vector of values for the cumulative distribution $P(X)$, corresponding to an input vector of random-variable values X_k. The function `qlnorm` returns a vector of values X_k for an input vector of values for the cumulative distribution $P(X)$.

a) For $\mu = 0$ and $\sigma = 1$, plot the log-normal PDF and CDF for $0 < X < 6$. Limit the y axis if there is a singularity in the PDF as $X \to 0+$.

b) For $\mu = 0$ and $\sigma = 1$, compute the expected value of the mean $\langle X \rangle = \int_0^\infty Xp(X)dX$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$.

c) For $\mu = 2$ and $\sigma = 1$, plot the log-normal PDF and CDF for $0 < X < 25$. Limit the y axis if there is a singularity in the PDF as $X \to 0+$.

d) For $\mu = 2$ and $\sigma = 1$, compute the expected value of the mean $\langle X \rangle = \int_0^\infty Xp(X)dX$ with a numerical integral over an appropriate range of X with at least 1000 values of X and $p(X)$.

Problem 4: R Functions for Binomial Distribution

The binomial probability density function is a discrete PDF in the instances of a specific event, such as "heads" in a coin flip, with fixed probability p follows a sequence of probabilities after N discrete trials. The random variable K is the number of instances of the specific event that occur within a set of N trials. The variable K is an integer in the range $[0, N]$. This PDF applies for K as

$$p(K) = \binom{N}{K} p^K (1-p)^{(N-K)}$$

where $\binom{N}{K}$ is the number of combinations of K items out of N items without replacement, that is

$$\binom{N}{K} = \frac{N!}{K!(N-K)!}$$

The log-normal PDF is represented by four R functions. The function `rbinom` returns a vector of random-variable realizations of $p(X)$, as we saw in Problem Set 1. The function `dbinom` returns a vector of $p(X)$ values, corresponding to an input vector of random-variable values X_k. The function `pbinom` returns a vector of values for the cumulative distribution $P(X)$, corresponding to an input vector of random-variable values X_k. The function `qbinom` returns a vector of values X_k for an input vector of values for the cumulative distribution $P(X)$.

a) For $N = 10$, $p = 0.5$, plot the binomial PDF and CDF for $0 \leq K \leq 10$. For $N = 100$, $p = 0.5$, plot the binomial PDF and CDF for $0 \leq K \leq 100$.

b) For $N = 10$, $p = 0.1$, plot the binomial PDF and CDF for $0 \leq K \leq 10$. For $N = 100$, $p = 0.1$, plot the binomial PDF and CDF for $0 \leq K \leq 100$.

c) Assume that great earthquakes (Richter magnitude $M \geq 8.0$) occur yearly in a plate-boundary segment, say the Western Aleutians, with a fixed probability $p = 0.01$ events per year. Assume also that each year's earthquakes are statistically independent of all other years' earthquakes. In a single century, what is the probability of $K \geq 3$ earthquakes on this plate-boundary segment? In a single century, what is the probability of $K \geq 5$ earthquakes on this plate-boundary segment?

d) Let p_3 be the probability of $K \geq 3$ earthquakes on the plate-boundary segment in part (c). The typical number of centuries that one might expect to occur in order to observe a single instance of $K \geq 3$ great earthquakes is the reciprocal of this probability, that is, $M_3 \approx (p_3)^{-1}$

centuries. Use the function `rbinom` to realize 100 centuries of great-earthquake activity with these binomial statistics. How many centuries in your binomial realizations have $K \geq 3$ events? Do these realizations comport with the probability you estimated in part (c)?

Problem 5: Bayes' Theorem and Rapid COVID-19 Tests

The use of Bayes' Theorem is important for rapid COVID-19 testing, because the probabilities of false positives and false negatives can be substantial. Let A be the set of all people who, on the day of testing, are infected with COVID-19. The complement of this set A^C is the set of all people who are not infected with COVID-19. These sets satisfy the probability that $P(A^C) = 1 - P(A)$, that is, a person has unit probability of being in either A or A^C and zero probability of being in both.

If a population takes a rapid test for COVID-19, the set of all positive tests is B, the set of all negative tests is B^C, and the probability of a positive test is $P(B)$. Some of the positive tests will be accurate, with conditional probability $P(B|A)$, the probability of the person being a member of B (testing positive) if they are a member of A (are infected). The conditional probability of a false positive is $P(B|A^C)$. We assume that these conditional probabilities are known, perhaps by testing patients with both a rapid COVID test and with a more-accurate, but slower, PCR COVID test. The conditional probability of a false negative, that is, of an infected person testing negative for COVID with the rapid test, is $P(B^C|A)$. We want to know the probability $P(A|B)$ that a person is infected, that is, is a member of set A, if they test positive, that is, is a member of B. From Bayes' Theorem,

$$P(A|B) = \frac{P(A)P(B|A)}{(P(A)P(B|A) + P(A^C)P(B|A^C))}$$

where the denominator is

$$(P(A)P(B|A) + P(A^C)P(B|A^C)) = P(B)$$

Consider the performance of an at-home COVID test that has recently been authorized for emergency use by the FDA.[1] The Abbot BinaxNow test is 84% accurate in detecting positive cases and 98% accurate in detecting negative cases. This translates into $P(B|A) = 0.84$ and $P(B|A^C) = 0.02$.

(a) If the background prevalence of COVID-19 infection in the general population is 10%, that is, if $P(A) = 0.10$, what is the probability that a positive test on the Abbot BinaxNow at-home test indicates a COVID infection, and not a false positive?

(b) If the background prevalence of COVID-19 infection in the general population is 2%, that is, if $P(A) = 0.02$, what is the probability that a positive test on the Abbot BinaxNow at-home test indicates a COVID infection, and not a false positive?

(c) Suppose 1000 randomly selected people take the Abbot BinaxNow at-home test and 100 receive a positive test for COVID. By manipulating the formulas above, determine the likely probability $P(A)$ for COVID-19 in the general population.

(d) Suppose 1000 randomly selected people take the Abbot BinaxNow at-home test and 20 receive a positive test for COVID. By manipulating the formulas above, determine the likely probability $P(A)$ for COVID-19 in the general population.

[1] https://www.wcnc.com/article/news/verify/verify-yes-at-home-covid-19-tests-are-reliable-to-determine-whether-youve-been-infected/275-89e83e06-f7c1-4b1b-bb78-d8cea92dc906