Probability and Statistics
for Final Year Engineering Students

By Yoni Nazarathy, Last Updated: May 24, 2011.

Exercises and Tutorial 2:
Independence and Sampling Distributions

Independence:

Two random variables X, and Y are independent if $P(X \in A \ \cap Y \in B) = P(X \in A)P(X \in B)$.

1. For each of the following cases, indicate if the assumption of independence sensible:
    a. X is the result of a die throw and Y is the result of a coin flip.
    b. X is the result of a die throw and Y takes 0 if the result is even and 1 if the result is odd.
    c. X is the amount of rainfall in day i and Y is the amount of rainfall in day i+1.
    d. X is the amount of rainfall in day i and Y is the amount of rainfall in day i+100.
2. Calculate the probability of getting the sequence (H,H,T,H,T) in 5 independent coin flips (with probability of H being ½):
    a. Using the independence property.
    b. By counting.
3. Modify the previous problem by assuming that in the n'th coin flip, the probability of H is 1/n. What is now the probability of the sequence (H,H,T,H,T)? Can the problem still be solved by means of counting?
4. You generate two independent uniformly distributed random variables in the range [0,1], U_1 and U_2. You let I_1 be 1 if $U_1 < 2/3$ and 0 otherwise. Similarly I_2 is 1 if $U_2 < 2/3$ and 0 otherwise.
    a. What is the CDF of I_1?
    b. What is the expected value of I_1?
    c. What is the probability that $Y = I_1 * I_2$=1 ?
    d. What is the CDF of Y?
    e. Is it true that $P(U_1 < \frac{1}{4}, I_1 = 1) = P\left(U_1 < \frac{1}{4}\right) P(I_1 = 1)$?
5. You generate a sequence of n independent uniformly distributed random variables: $U_1, U_2, \ldots, U_n$. The random variable U_i is distributed on the range [0,i]. (I.e. U_1 takes values between 0 and 1, U_2 takes values between 0 and 2, etc…). What is the probability that all of the random variables are less than 1?
6. Similarly to the previous exercise. You generate a sequence of n independent random variables where each variable has a continuous distribution taking values in the range $[0, \infty)$ and the i'th random variable has density $f_i(x) = i\, e^{-i\, x}$. What is the probability that all of the random variables are greater than 1?
7. Consider a coin flip let I_1 be 1 if heads and 0 otherwise. Let I_2 be 1 if tails and 0 otherwise. Are these independent random variables?

Sampling with replacement (independent trials):

n – the number of samples.

p – the probability of "success".

X – the number of successes (can take values 0,…,n).

$X \sim Binomial(n,p)$, $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}, E[X]=np, Var(X)=n\,p(1-p)$

8. You decide to completely guess on a multiple-choice test that has 17 questions each question with 4 answers. What is the probability of getting a grade greater or equal to 50%.
9. A short communication message contains 32 bits. Bit values are assumed to be independent. The proportion of 1's is 1/3 and the proportion of 0's is 2/3.
    a. Write an expression for the probability of having all 1's.
    b. Write an expression for the probability of having all 0's.
    c. Write an expression for the probability of having a single 1 and the rest 0's.
    d. What is the expected value of the sum of the bits?
    e. Write an expression for the probability of having 3 0's and the rest 1's.
10. A coin having a probability of heads being 0.4 is tossed 5 times. What is the probability of obtaining an even number of heads.
11. A container has room for exactly 5 boxes which can either weigh 2 tons each or 3 tons each. Containers are being packed by allocating boxes at random, where the proportion of 2 ton boxes is 30% and 3 ton boxes are 70%. What is the mean container weight? Draw the probability mass function of the container weight. What proportions of containers weigh more than 11 tons?
12. Car tune-up times in a garage are assumed to be independent and to have a distribution with density $f(x)=2\,e^{-2x}$. What is the probability that out of 10 tune ups, more than 7 tune ups took a duration longer than 1 time unit?
13. Let X_1 and X_2 be two independent binomial random variables with parameters (n_1,p_1) and (n_2,p_2) respectively.
    a. What is $E[X_1 + X_2]$?
    b. What is Var$(X_1 + X_2)$?
    c. In case where $p_1=p_2=p$. Write the PDF of $X_1 + X_2$.
14. The proportion of "marked items" in a population is p. You use a random sample of size n=3 to estimate the proportion, obtaining $\hat{p}$. What is the CDF of $\hat{p}$?

Sampling without replacement:

N – the number of items of type 1 in the population.
M – the number of items of type 2 in the population.
n – the number of samples taken.
X – the number of "successes", items of type 1.

$X \sim Hypergeometric(n, N, M)$, $P(X = k) = \frac{\binom{N}{k}\binom{M}{n-k}}{\binom{N+M}{n}}$, $E[X] = n\frac{N}{N+M}$.

The range of values that X can take are,

$$\max(0, n - M) \leq k \leq \min(n, N)\,.$$

15. A fish pond has 12 white fish and 18 gold fish. Five fish are taken out at random without replacement. What is the probability that 3 of them are white?
16. Repeat the previous exercise assuming that after taking a fish, it is returned to the pond.
17. Repeat the previous two exercises assuming 120 white fish and 180 gold fish and checking the probability that 30 of them are white. How do the answers differ?
18. Give numeric examples of hypergeometric distributions that takes values in the following ranges:
    a. 0,…,n
    b. n-M,….,N
    c. 0,…,N
    d. n-M,…,n
19. A new neighborhood has 7 square lots and 4 trapezoidal lots. Lots are given out to people by a lottery system in a completely random manner. 5 families apply for lots. What is the probability that a single trapezoidal lot is not taken (the other trapezoidal lots have been allocated to families)?

The Gaussian Distribution:

μ - the mean.

σ – the standard deviation.

X – a random quantity.

$X \sim Normal(\mu, \sigma^2)$, $F(x) = P(X \le x) = \int_{-\infty}^{x} \frac{1}{\sigma} f\left(\frac{u-\mu}{\sigma}\right) du$, where $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$.

The function $F(x)$ can not be evaluated explicitly and requires numeric integration. The values of $F(x)$ appear in a normal distribution table. Some calculators give $F(x)$ (in TI calculators this is the normCDF function under the dist menu). Here is a normal distribution table:

	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

.

Observe that F(0)=1/2. Why? And that $F(3) \cong 1$ over 99% of the area under the normal curve lies between -3 and 3.

20. Let $Z \sim Normal(0,1)$, find
    a. $P(Z \le 1)$
    b. $P(Z \le 1.55)$
    c. $P(Z \le 1.557)$
    d. $P(Z \le -1)$
    e. $P(-1 \le Z \le 1)$
    f. $P(Z \ge 2.3)$

21. Let $X \sim Normal(-20, 2.3^2)$, find
    a. $P(X \leq -12)$
    b. $P(X \leq 0)$
    c. $P(Z \leq -1)$
    d. $P(-22.3 \leq X \leq -17.7)$
    e. $P(X \geq -21.2)$
22. Given the probability γ, find the *percentile* x, such that $P(X \leq x) = \gamma$.
    a. X is a standard normal random variable and $\gamma = 0.9984$.
    b. X is a standard normal random variable and $\gamma = 0.5$.
    c. $X \sim Normal(10, .35^2)$ and $\gamma = 0.2$.
23. $Let\ X \sim Normal(\mu, \sigma^2)$. Find:
    a. $P(\mu - \sigma \leq X \leq \mu + \sigma)$
    b. $P(\mu - 2\sigma \leq X \leq \mu + 2\sigma)$
    c. $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma)$
24. Let $X_1, X_2, \ldots, X_n$ be a sequence of independent random variables with the same mean μ for all random variables and $\sqrt{Var(X_i)} = \sigma_i$.
    a. Assume $\sigma_i = \sigma$ (constant) for all i. What is the probability that all random variables are greater than $\mu + \sigma$?
    b. Assume $\sigma_i = \sigma$ (constant) for all i and let n=20. Calculate the probability that 12 of the random variables are greater than $\mu + \sigma$ (and 8 of them are less than $\mu + \sigma$).
    c. Let $\sigma_i = \frac{1}{\sqrt{i}}$ and let n=3. What is he probability that all 3 random variables are greater than $\mu + 1$?

The Central Limit Theorem:

Let $X_1, X_2, \ldots$ be a sequence of independent random variables having the same distribution with mean μ and variance σ^2. Then:

I. $Y_n = \sum_{i=1}^{n} X_i$ is asymptotically normally distributed with mean $n\mu$ and variance nσ^2.

II. Alternatively, the sample mean $\bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$ is asymptotically normally distributed with mean μ and variance $\left(\frac{\sigma}{\sqrt{n}}\right)^2$.

III. Alternatively, there is the case where $X_1, X_2, \ldots$ is a Bernoulli (binary) sequence with success probability p, denoted $I_1, I_2, \ldots$. Then $E[I_i] = p$ and $Var(I_i) = p(1-p)$ and $B_n = \sum_{i=1}^{n} X_i$is a Binomial(n,p) random variable. Then following (I), B_n is asymptotically normally distributed with mean np and variance $np(1-p)$.

IV. Alternatively, the sample proportion: $\hat{p}_n = \frac{\sum_{i=1}^{n} I_i}{n}$ is asymptotically normally distributed with mean p and variance $\left(\frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)^2$.

25. Observe the central limit by doing a simple Excel (or similar) simulation:
    a. First generate random variables uniformly distributed over the range [0,1]. E.g. In Excel create 5 columns, each containing 10,000 such random variables.
    b. Add the random variables to obtain 10,000 copies of $\bar{X}_5$.
    c. Calculate the sample mean and sample standard deviation of the result, plot the histogram.
    d. Compare Q3 of the resulting values (this the 7,500'th observation when sorting the 10,000 samples of $\bar{X}_5$) to the 0.75'th percentile calculated from the appropriate normal distribution.

26. A passenger jet is designed to carry up to 200 passengers each having luggage of no more than 35Kg. Studies have shown that the actual distribution of luggage that a passenger carries can be approximated by a distribution (which is not normal) but has mean 34 Kg and a standard deviation of 7 Kg. Approximate the probability that the jet carries more than 7200 Kg of luggage.

27. Suppose that the proportion of defective items in a large manufactured lot is 0.2. What is the smallest random sample of items that needs to be taken from the lot in order for the probability to be at least 0.97 that the proportion of defective items in the sample will be less than 0.25?

Selected Solutions

1) a: Yes, b: No, c: Typically No, d: Typically Yes.

2) a: $\frac{1}{2}\frac{1}{2}(1-\frac{1}{2})\frac{1}{2}(1-\frac{1}{2})=\frac{1}{32}$ b: $\frac{number\ outcomes\ yielding\ HHTHT}{total\ number\ of\ outcomes}=\frac{1}{2^5}$

3) $\frac{1}{1}\frac{1}{2}(1-\frac{1}{3})\frac{1}{4}(1-\frac{1}{5})=\frac{1}{15}$. No, counting does not work because there are different probabilities for different outcomes (this is not a symmetric probability space).

4) a) $F(x)=\begin{cases}0 & x<0\\ \frac{1}{3} & 0\le x<1\\ 1 & 1\le x\end{cases}$ b) 2/3 c) $P(Y=1)=P(both\ are\ 1)=\frac{2}{3}\frac{2}{3}=\frac{4}{9}$

d) Y is gets 1 w.p. 4/9 and 0 w.p 5/9, so: $F(x)=\begin{cases}0 & x<0\\ \frac{5}{9} & 0\le x<1.\\ 1 & 1\le x\end{cases}$ E) No.

5) $\frac{1}{n!}$.

6) $p(X_i>1)=\int_1^\infty i\,e^{-i\,x}dx=e^{-i}$.

P$(X_1>1,\dots,X_n>1)=e^{-1}e^{-2}\dots e^{-n}=e^{-(1+\cdots+n)}=e^{-\frac{n(n+1)}{2}}$

7) No: $P(I_1=1,I_2=1)\neq P(I_1=1)P(I_2=1)$.

8) $X\sim Bin(17,\frac{1}{4})$. $P(pass)=P\left(X\ge\frac{17}{2}\right)=P(X\ge 9)=\sum_{k=9}^{17}\binom{17}{k}\frac{1}{4}^k\frac{3}{4}^{17-k}$.

9) $a)\ \left(\frac{1}{3}\right)^{32}$ b) $\left(\frac{2}{3}\right)^{32}$ c) $32\frac{1}{3}\left(\frac{2}{3}\right)^{31}$ d)$\frac{32}{3}$ e)$\frac{32*31*30}{6}\left(\frac{2}{3}\right)^3\left(\frac{1}{3}\right)^{29}$

10) $X\sim Bin(5,0.4)$

$P(X=even)=P(X=0)+P(X=2)+P(X=4)=0.6^5+10\ *\ 0.4^2*0.6^3+5*\ 0.4^4*0.6=0.50016$

11) Let X be the number 3 ton boxes. Weight W=3*X+2*(5-X)=10+X.
E[W]=E[10+X]=10+E[X]=10+5*0.7=13.5.
The PDF of W looks similar to that of Bin(5,0.7) but shifted 10 units to the right.
P(W>11)=1-P(W<=11)=1-P(W=11)-P(W=10)=1-P(X=1)-P(X=0) = …

12) X=Number of tune ups longer than 1 time unit. X~Bin(10,p) with p=$\int_1^\infty f(x)dx=e^{-2}=0.1353$.
Now calculate P(X>7)=P(X=8)+P(X=9)+P(X=10).

15) X=Number of white fish X~HG(5,12,18). P(X=3)=0.236201Type equation here.

16) Now use a binomial distribution, X~Bin(5,12/30). P(X=3)=0.2304

17) Now the answers will be very close to each other.

20) a) 0.8413 d) P(Z<=-1)=1-P(Z<=1)=0.1587 Type equation here.
e) P(-1<=Z<=1) = F(1)-F(-1)=F(1)-(1-F(1))=2F(1)-1=0.6826

21)

d) $P(-22.3\le X\le -17.7)=P\left(\frac{-22.3-(-20)}{2.3}\le\frac{X-(-20)}{2.3}\le\frac{-17.7-(-20)}{2.3}\right)=P(-1\le Z\le 1)=0.6826.$

22) a) Look at the normal table and find the point 0.9984 inside the table. $z_{0.9984} = 2.95$.

b) 0.

c) $0.35\ z_{0.2} + 10 = 0.35(-z_{0.8}) + 10 = -0.35 * 0.845 + 10 = 9.70425$.

23) a) 0.6823 b) 0.9545 c) 0.9973

26) $E[X_i] = 34, Var(X_i) = 7^2$. $Y = \sum_{i=1}^{200} X_i$. $E[Y] = 6800, Var(Y) = 98.99^2$.

According to the CLT (and assuming that passenger weights are independent), Y is approximately normally distributed. So, $P(Y > 7200) = P(Z > 4.04) \approx 0$.

27) We know that p=0.2 and that $n\,\hat{p} \sim Bin(n, 0.2) \sim Normal\left(n0.2, \left(0.4\sqrt{n}\right)^2\right)$.

$$P(\hat{p} < 0.25) = P(n\hat{p} < n0.25) = P\left(Z < \frac{n0.25 - n0.2}{0.4\sqrt{n}}\right) < 0.97$$

So solve: $\frac{n0.25-n0.2}{0.4\sqrt{n}} = z_{0.97} = 1.89$ and get n=228 so we need n=230.