Stat 110

Notes by Edward Chen

October 2022

Preface

These are my notes to Harvard's Stat 110 class, taught by Joseph Blitzstein and Jessica Hwang. The class covers all the basics of probability–counting principles, probabilistic events, random variables, distributions, conditional probability, expectation, and Bayesian inference.

Materials were collected from Blitzstein's book and lecture videos.

1 Probability

Probability gives us a logical framework to view and analyze uncertainty. It is the foundation and language for statistic ass well as the basis for topics such as Statistics, Physics, Biology, and Computer science.

Definition 1.1. A **sample space** S is the set of all theoretically possible outcomes.

Let D be the set of all coin flips with at least two consecutive heads. The sample space, expressed as a set would be:

$$D = \cup_{j=1}^{9}(A_j \cap A_{j+1})$$

Definition 1.2. Some more set notation:

- sample space: S
- s is a possible outcome: $s \in S$
- A is an event: $A17S$
- A implies B:
- A and B are mutually exclusive: $A \cap B = \emptyset$

Example 1.3. (Birthday Problem). There are k people in a room. Assume that each person's birthday is equally likely to be any of the 365 days of the year and that they are independent. What is the probability that at least one pair of people in the group have the same birthday?

Solution

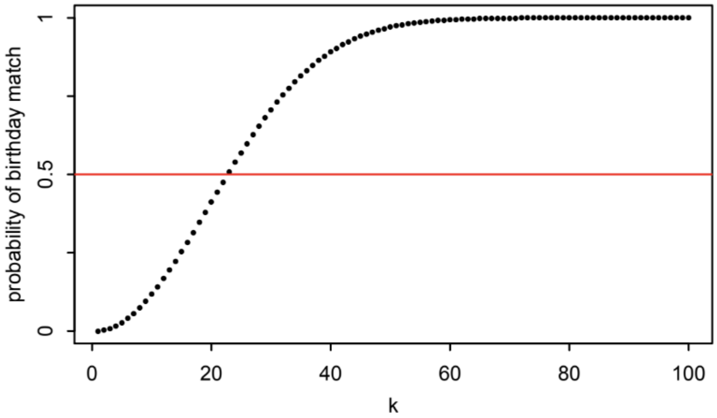
There are 365^k ways to assign birthday as for each of the k people, one of the 365 days is chosen. Counting directly here is difficult, but it becomes much easier when we count the complement: the number of ways to assign birthdays to k people such that no two people share a birthday:

$$P(\text{no birthday match}) = \frac{365 \cdot 354 \cdots (365 - k + 1)}{365^k}$$

and thus, we get:

$$P(\text{at least 1 birthday match}) = 1 - \frac{365 \cdot 354 \cdots (365 - k + 1)}{365^k}$$

And plotted,



Theorem 1.4. *(Binomial coefficient formula). For $k \leq n$, we have:*

$$\binom{n}{k} = \frac{n(n-1)...(n-k+1)}{k!} = \frac{n!}{(n-k)!k!}$$

Some proof methods covered in the chapter include complementary counting, stars and bars, and story proofs(1.5).

There are a lot of formulas with binomial coefficients:

$$\binom{n}{k} = \binom{n-k}{k}$$

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^k y^{n-k}$$

$$n\binom{n-1}{k-1} = k\binom{n}{k}$$

$$\text{Vandermonde's}: \binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}$$

Example 1.5. (Partnerships). Let's prove

$$\frac{(2n)!}{2^n \cdot n!} = (2n-1)(2n-3)...3 \cdot 1$$

Story proof:: We will show that both sides count the number of ways to break $2n$ people into n partnerships. Take $2n$ people, and give them ID numbers from 1 to $2n$. We can form partnerships by lining up the people in some order and then saying the first two are a pair, the next two are a pair, etc. This over counts by a factor of $n! \cdot 2^n$ since the order of pairs doesn't matter, nor does the order within each pair. Alternatively, count the number of possibilities by noting that there are $2n-1$ choices for the partner of person 1, then $2n-3$ choices for person 2, and so on.

Definition 1.6. (General definition of probability). A *probability space* consists of a sample space S and a *probability function* P and returns a real number between 0 and 1, $P(A)$, where A is the event it takes in.

Unlike the naive definition, here we can have events with different probabilities.

The *frequentist* view of probability is that it represents a long-run frequency over a large number of repetitions of an experiment. The *Bayesian view* is that it represents a degree of belief about the event in question, so we can assign probabilities to hypotheses.

Inclusion-exclusion example: With a triple venn-diagram, we can write:

$$P(A \cup B \cup C) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cup B \cup C)$$

Generally, we can write:

Example 1.7. (de Montmort's matching problem). Consider a shuffled deck of n cards, from 1 to n. You flip each one over one by one. What is the probability the ith index card you turn over has value i?

With PIE, we get

$$P(\cup_{i=1}^{n} A_i) = \frac{n}{n} - \frac{\binom{n}{2}}{n(n-1)} + \frac{\binom{n}{3}}{n(n-1)(n-2)} - ... + (-1)^{n+1} \cdot \frac{1}{n!}$$

$$= 1 - \frac{1}{2!} + \frac{1}{3!} - ... + (-1)^{n+1} \cdot \frac{1}{n!}$$

With large n, this approaches the Taylor series for $\frac{1}{e}$:

$$e^{-1} = 1 - \frac{1}{1!} + \frac{1}{2!} - \frac{1}{3!} + ...$$

R: Please read section 1.8 for an introduction to R. R allows us to simulate and deal with large sets of data.

2 Chapter 2: Conditional Probability

Conditional Probability explores how we should update our probability when we receive new information.

Definition 2.1. If A and B are events with $P(B) > 0$, then conditional probability can be expressed as:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Example 2.2. (Two Children): Martin Gardner posed the following puzzle:

1. *Mr. Jones has two children. The older child is a girl. What is the probability that both children are girls?*

2. *Mr. Smith has two children. At least one of them is a boy. What is the probability that both children are boys?*

The intuition is that they should both be $\frac{1}{2}$, but the respective probabilities are actually $\frac{1}{2}$ and $\frac{1}{3}$. Example 2 showcases conditional probability: we know that at least one of them is a boy so there are only three possibilities: BG, GB, BB.

Theorem 2.3. *(Bayes' rule).*

$$P(A|B) = \frac{P(A \cap B)}{B} = \frac{P(B|A)P(A)}{P(B)}$$

Definition 2.4. (Odds). The *odds* of an event A are

$$\text{odds(A)} = P(A)/P(A^c)$$

. Ex.: if $P(A) = 2/3$, we say the odds in favor of A are 2 to 1.

Example 2.5. (Unanimous agreement). The article "Why too much evidence can be a bad thing" by Lisa Zyga says:

> Under ancient Jewish law, if a suspect on trial was unanimously found guilty by all judges, then the suspect was acquitted. This reasoning sounds counter-intuitive, but the legislators of the time had noticed that unanimous agreement often indicates the presence of systemic error in the judicial process.

This is because a systemic error occurs independent of the conviction, so there may be an unseen bias at play here.

Definition 2.6. (Independence) Events A and B are *independent* if $P(A \cap B) = P(A)P(B)$.

For infinitely many events, we say that they are independent if every finite subset of the events is independent.

Definition 2.7. (Conditional independence). Events A and B are said to be *conditionally independent* given E if $P(A \cap B|E) = P(A|E)P(B|E)$.

> It is easy to make terrible blunders stemming from confusing independence and conditional independence. Two events can be conditionally independence given E, but not independent given E^c. Two events can be conditionally independent given E, but not independent. Two events can be independent, but not conditionally independent given E. Great care is needed

Definition 2.8. (Coherence) An important property of Bayes' rule is that is is *coherent*: for the same updates, no matter the update size, the final state should be the same.

For example, a person take two tests for an infectious disease. Given that the first test is positive and the second test is negative, the posterior probability between updating using Bayes' rule both tests at a time and updating the first test and then the second test is identical.

Example 2.9. (Monty Hall):

Here the strategy is to use conditioning as a problem-solving tool. If we knew where the car actually is, it makes the probabilities a lot easier to deal with:

$P(\text{get car}) = P(\text{get car}|C_1) \cdot \frac{1}{3} + P(\text{get car}|C_2) \cdot \frac{1}{3} + P(\text{get car}|C_3) \cdot \frac{1}{3} = \frac{2}{3}$

That is why it is always better to switch because now you have a 2/3 success strategy.

Example 2.10. (Gambler's ruin):

The gambler's ruin problem is a problem where we have two players, one with n dollars and the other with n dollars. They play a game where they bet 1 dollar each. There is probability of p that the first player wins and gets 2 dollars, and otherwise the second player gets the 2 dollars. The game ends when one of the players has 0 dollars. What is the probability that the first player wins?

This can be solved by conditioning on the first step and is just a general version of the random walk problem. We have **absorbing state** at 0 and $2n$ and we have the relationship:

$$p_i = p \cdot p_{i+1} + (1-p) \cdot p_{i-1}$$

To solve, we start by guessing $p_i = x^i$. Then we get:

$$x^i = p \cdot x^{i+1} + (1-p) \cdot x^{i-1}$$

$$px^2 - x + (1-p) = 0$$

$$x = \frac{1 \pm \sqrt{1 - 4p(1-p)}}{2p}$$

3 Random Variables

In this course, we look at a **random variable** as more than ”a variable that takes on random values.” A better definition is that a random variable is a function that maps a sample space to the real numbers.

Definition 3.1. (Bernoulli distribution). The **Bernoulli distribution** is the probability distribution of a random variable X that takes on the values 0 and 1 with probabilities p and $1 - p$, respectively. We write X $Bern(n, p)$ as X is distributed as a Bernoulli function with n independent Bernoulli random variables that have probability p of being 1.

Interestingly: the binomial distribution is the same distribution of the sum of n independent Bernoulli random variables.

Todo: proposition 9.2

Definition 3.2. (Probability mass function). The **probability mass function** of a discrete random variable X is a function f_X that assigns to each value x of X the probability that X takes on the value x.

Definition 3.3. (Cumulative distribution function). The **cumulative distribution function** of a discrete random variable X is a function F_X that assigns to each value x of X the probability that X takes on a value less than or equal to x.

Definition 3.4. (Hyper-geometric distribution) is the probability distribution of the number of successes in a sample of size n drawn with replacement from a finite population of size N containing exactly K successes.

Theorem 3.5. *(Hyper-geometric PMF): The probability mass function of a hyper-geometric random variable X is given by:*

$$f_X(x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}$$

Hyper-geometric distributions are very similar to binomial distributions in that they take on n Bernoulli trials, but note that hyper-geometric distributions are dependent as we are drawing without replacement.

4 Expectation

Here is a fundamental derivation of binomial expectation for $X \sim Binom(n,p)$:

$$EV(X) = \sum_{k=0}^{n} k\binom{n}{k}p^k q^{n-k} = n\sum_{k=0}\binom{n-1}{k-1}p^k q^{n-k} = np\sum_{j=0}\binom{n-1}{k-1}p^j q^{n-1-j} = np$$

Geometric and Negative Binomial Distributions

Definition 4.1. (Geometric distribution). The **geometric distribution** is the probability distribution of the number of Bernoulli trials needed to get the first success.

Definition 4.2. (Negative binomial distribution). The **negative binomial distribution** is the probability distribution of the number of Bernoulli trials needed to get r successes.

Negative binomial expectation is just given by: $E(x) = E(x_1) + E(x_2) + ... + E(x_r) = r \cdot \frac{q}{p}$.

Poisson Distribution

Poisson Distribution is a common distribution used for modeling the number of "successes" given a low-probability, high frequency event. The PMF is given a parameter λ by:

$$P(x = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

It is a valid PMF as $\sum_{k=0}^{\infty} P(x = k) = e^{-\lambda} \cdot e^{\lambda} = 1$. With the same derivation as above, we can show that the expected value of a Poisson distribution is λ.

Connection with binomial distribution: actually converges pretty well for large n, small p_j, but here events can also be "weakly dependent," and have different p_j's.

Poisson can be used for finding the approximate probability function of the birthday problem, where lets say we have n people and want to find probability of there is exactly k triplets of people with the same birthday. We know that $EV = \lambda = \binom{n}{3}(\frac{1}{365})^2$ so we can create a Poisson distribution. Notice there is some weak dependence here.

5 Continuous Distributions

We've learned about DPFs, but now we want can say similar things about **probability density functions**, or PDFs. In fact, to they are easily using the Fundamental Theorem of Calculus in that to get a PDF is just to take the derivative of DPF's.

Variance

Definition 5.1. (Variance). The **variance** of a random variable X is defined as:

$$Var(X) = E[(X - EX)^2] = Ex^2 - (Ex)^2$$

Something more interpretable is standard deviation: $\sigma = \sqrt{Var(X)}$.

But how do we actually find Ex^2? We can use the Fundamental Theorem of Calculus to get:

$$Ex^2 = \int_{-\infty}^{\infty} x^2 f_X(x) dx$$

Generally, this is called the Law of the Unconscious Statistician(LOTUS):

$$E(g(x)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Uniform and Universality of the Uniform

Given being able to produce random variables from a uniform distribution, we can produce any other distribution.

The proof follows that if we let F be the CDF of X, then $F^{-1}(u) \sim F$ if $U \sim$ Unif(0,1). The intuition is that we plug a random variable into its own CDF.

Normal Distribution

The normal distribution is a very important distribution that is used to model many things. It is defined by two parameters, μ and σ^2. The PDF is given by:

$$f_X(x) = ce^{-\frac{z^2}{2}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Central Limit Theorem: the sum of independent random variables tends to a normal distribution.

Main intuition here is that we can actually integrate the definite integral of $ce^{-\frac{z^2}{2}}$ using polar coordinates to get $\frac{1}{\sqrt{2\pi}}$.

6 Moments

The nth *moment* of an r.v. X is $E(X^n)$. In this chapter, we xplore how the moments of an r.v. shed light on its distribution. We have already seen that the first two moments are useful since they provide the mean $E(X)$ and the variance $E(X^2) - (EX)^2$.

Theorem 6.1. *Let X be an r.v. with mean μ, and let m be a median of X.*

- *The value of c that minimizes the mean squared error $E(X-c)^2$ is $c = \mu$.*
- *A value of c that minimizes the mean absolute error E—X-c— is $c = m$.*

Proof.

We first prove that $E(X-c)^2 = Var(X) + (\mu - c)^2.$, given from $Var(X) = Var(X-c) = E(X-c)^2 - (E(X-c))^2 = E(X-c)^2 - (\mu-c)^2$. Thus, mean squared error is minimized at $c = \mu$.

Next, for mean absolute error, note that for $X \leq m$, $|X-a|-|X-m| = a-m$ and $|X-a| - |X-m| = m-a$. Using indicator r.v. I for $X \leq m$ and $1-I$ for $X > m$, then

$$\begin{aligned} E(Y) &= E(YI) + E(Y(1-I)) \\ &\geq (a-m)E(I) + (m-a)E(1-I) \\ &= (a-m)P(X \leq m) + (m-a)P(X > m) \\ &= (a-m)P(X \leq m) - (a-m)(1-P(X \leq m)) \\ &= (a-m)(2P(X \leq m) - 1). \end{aligned} \tag{1}$$

The expression $(2P(X \leq m) - 1) \geq 0$ is exactly the definition of median and thus must minimize it!

Now, we describe moments of higher dimensions. To explain why these are necessary, take the following two different distributions in Figure 1. Notice that PDF's of the same mean and variance can describe wildly different distributions.

We use the third and fourth moments, which describe the *asymmetry* of a distribution and the behavior of the *tails* or extreme values as seen in Figure 2.

Let us formally define moments.

Definition 6.2. (Kinds of moments). Let X be an r.v. with mean μ and variance σ^2. For any positive integer n, the nth $standardizedmoment$ is $E((\frac{X-\mu}{\sigma})^n)$.

Definition 6.3. (Skewness). The *skewness* of an r.v. X with mean μ and variance σ^2 is the third standardized moment of X:

$$Skew(X) : E(\frac{X-\mu}{\sigma})^3$$

Proposition 6.4. (Odd central moments of a symmetric distribution). Let X be symmetric about its mean μ. Then for any odd number m, the mth central moment $E(X_\mu)^m$ is 0 if it exists.

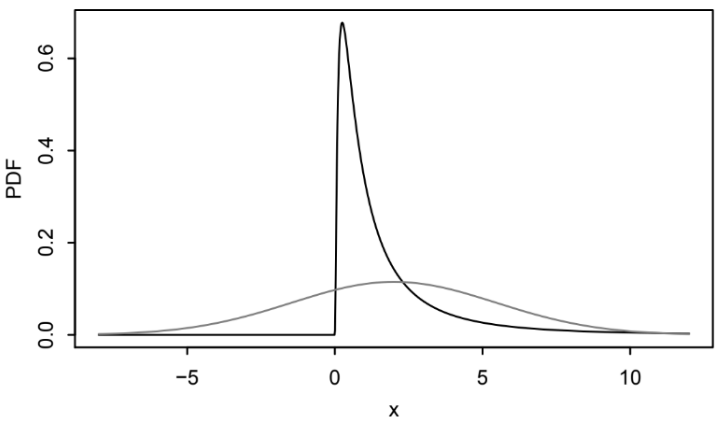


Figure 1: Two different distributions with mean=2 and variance=12. One is normal and the other is log-normal.

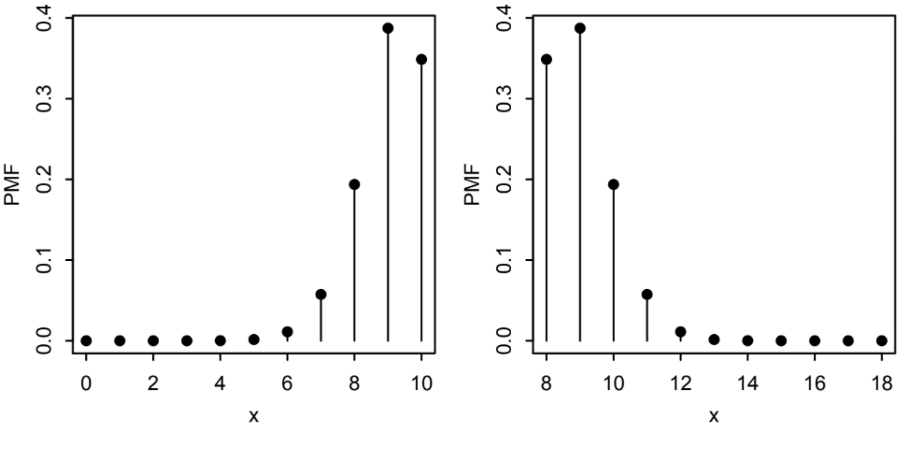


Figure 2: Left: Bin(10,0.9) is left-skewed. Right: Bin(10,0.1), shifted to the right by 8, is right-skewed but has the same mean, median, mode, and variance as Bin(10,0.9)

Using this proposition, calculating any odd moment is actually a good way to determine the skewness of a distribution. Most of the time though, using the third moment suffices and is easier to calculate.

Definition 6.5. (Kurtosis) The *kurtosis* of an r.v. X with mean μ and variance σ^2 is a shifted version of the fourth standardized moment of X:

$$Kurt(X) = E(\frac{X-\mu}{\sigma})^4 - 3$$

.

Notice that we subtract 3, making any Normal distribution have kurtosis 0.

The last thing I want to mention is that sampling moments can be an efficient way to determine the nth moment of a distribution. Given that the kth *sample moment* is the r.v.

$$M_k = \frac{1}{n}\Sigma_{j=1}^{n} X_j^k$$

Then, by the law of large numbers shows that the kth sample moment of i.i.d random variables $X_1, \ldots, X_n$ converges to the kth moment $E(X_1^k)$ as $n \to \infty$

6.1 Moment generating functions

Generating functions are useful here to bridge between sequence of numbers and the world of calculus. Given we start with a sequence of numbers, we attempt to create a continuous function that encodes the sequence.

Definition 6.6. A random variable X has moment-generating function (MGF)

$$M(t) = E(e^{tX})$$

if $M(t)$ is bounded on some interval $(-\epsilon, \epsilon)$ about zero.

We observe that:

$$M_{X+Y} = E(e^{t(X+Y)}) = E(e^{tX})E(e^{tY}) = M_X M_Y$$

The second inequality comes from the claim that if for X, Y independent, $E(XY) = E(X)E(Y)$

7 Joint distributions

When we first introduced random variables, we assumed that they were independent. However, this is not always the case. We can define a joint distribution as a distribution of two random variables. Just as univariate PMF's must sum to 1 and be nonnegative, the same must hold for joint PMF's:

$$\sum_x \sum_y P(X = x, Y = y) = 1$$

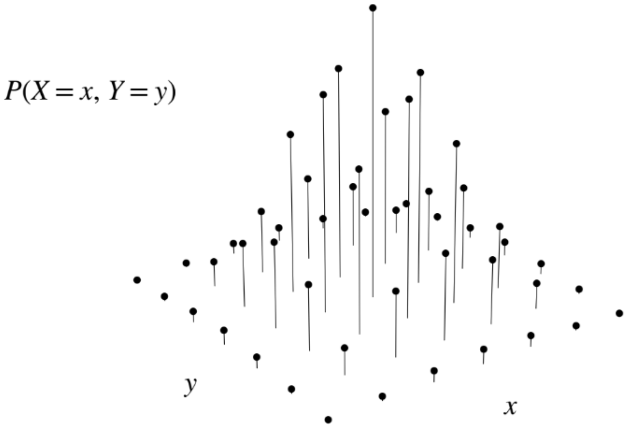


Figure 3: Joint distribution

Marginal PMF's, which describe the distribution of X ignoring Y, can be taken from joint PMF's. The marginal PMF $P(X = x)$ is obtained by summing over the joint PMF in the y-direction, as seen in Figure 4.

And normalizing the sum to 1 gives us a new PMF: conditional.

Dealing with continous PMF's becomes pretty simple given what we know about discrete.

Definition 7.1. (Joint PDF). If X and Y are continuous with joint CDF F_{X_Y}, their *joint* PDf is the derivative of the joint CDF with respect to x and y:

$$f_{X_Y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

Integrating, we can get the joint PDF of two r.v.s for any two-dimensional region.

$$P((X, Y) \in A) = \int \int_A f_{X,Y}(x, y) dx dy.$$

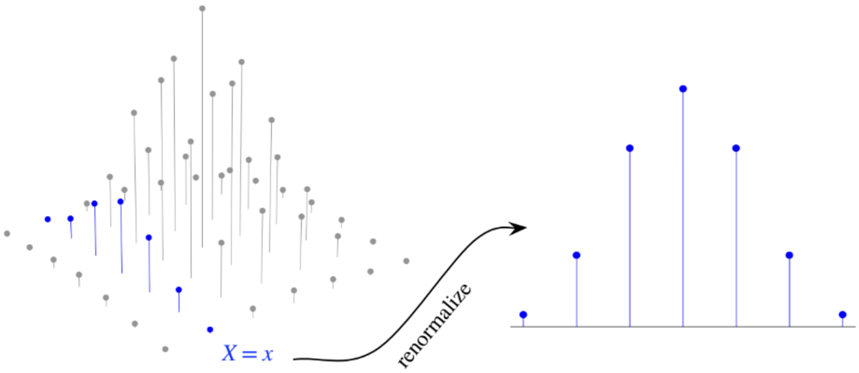


Figure 4: We can obtain the conditional PMF.

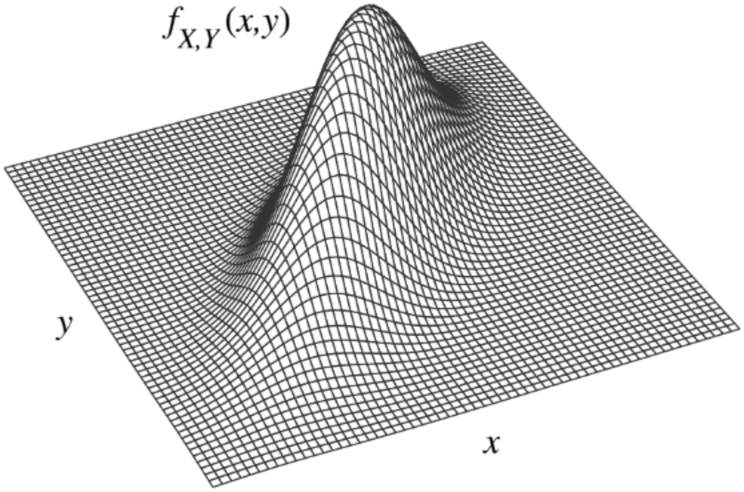


Figure 5: Joint PDF of continuous r.v.s X and Y.

7.1 Covariance and correlation

Covariance is a measure of how two random variables are related, while correlation is a normalized version of covariance.

Definition 7.2. (Covariance) The covariance between r.v.s X and Y is:

$$Cov(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = E[XY] - \mu_X\mu_Y$$

Definition 7.3. (Correlation)

$$Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

Theorem 7.4. *If X and Y are independent, then $Cov(X,Y) = 0$.*

Proof. We'll use the fact that $E[XY] = E[X]E[Y]$.

$$Cov(X,Y) = E[XY] - \mu_X \mu_Y = E[X]E[Y] - \mu_X \mu_Y = 0$$