Recitation Notes for PSTAT 120B

Haosheng Zhou

Jan, 2023

Week 1

This week we will review some of the important properties taught in 120A as a preparation for homework 0.

The first concept to review is the **continuous and discrete random variables**. Generally, a random variable

$$X(\omega):\Omega\to\mathbb{R} \tag{1}$$

is a mapping from the sample space to the real numbers, i.e. it assigns a value to each possible outcome of random experiment. Discrete random variables can take countably many values while continuous random variables can take uncountably many values. For example, if we want to consider the random variable X as the outcome after rolling one dice, then we have to first specify the **sample space**, i.e. the set of all possible outcomes rolling one dice, which should be $\Omega=\{1,2,...,6\}$. As a result, such random variable X is defined as

$$X(\omega)=\omega \tag{2}$$

an identity map. Since X can only take values in $\{1,2,...,6\}$, a finite set, it's a discrete random variable.

To describe a single random variable, we have the **cumulative distribution function (CDF)** defined for any random variable X as

$$F(x)=\mathbb{P}\left(X\leq x\right) \tag{3}$$

Such F is always right-continuous, increasing and $F(-\infty)=0, F(+\infty)=1$ (try to explain the meaning of those properties). In particular, for continuous random variable such F is continuous and for discrete random variable such F is a step function. For continuous random variables, assume that F is nice enough to be differentiable so $F'=f$ gives the **density** that characterizes the distribution of the continuous random variable (for random vectors, those concepts can be generalized).

To describe the relationship between two random variables, the most important property is **independence**. We call X,Y independent if

$$\forall x,y\in\mathbb{R}, \mathbb{P}\left(X\leq x, Y\leq y\right)=\mathbb{P}\left(X\leq x\right)\mathbb{P}\left(Y\leq y\right) \tag{4}$$

which can also be explained in the sense of conditional probability (try to write the equality in the conditional form). For discrete r.v. X,Y, they are independent if and only if $\forall x,y\in\mathbb{R}, \mathbb{P}\left(X=x,Y=y\right)=\mathbb{P}\left(X=x\right)\mathbb{P}\left(Y=y\right)$ and for continuous r.v. X,Y, they are independent if and only if $f_X(x)f_Y(y)=f_{X,Y}(x,y)$ *a.e.* (think about why the criterion for discrete r.v. does not hold for continuous r.v.).

The important concept to mention is the **expectation** of continuous or discrete random variables. For discrete random variable X, assume that its distribution is given by

$$p_k=\mathbb{P}\left(X=a_k\right)\ \ (k=0,1,...) \tag{5}$$

so the expectation is formed as

$$\mathbb{E}X = \sum_{k=0}^{\infty} a_k \cdot \mathbb{P}(X = a_k) = \sum_{k=0}^{\infty} a_k \cdot p_k \tag{6}$$

i.e., the **sum** of the product of the possible value a_k taken by X and the probability of X taking value a_k.

For discrete random variable X, assume that its density is $f(x)$, so the expectation is formed as

$$\mathbb{E}X = \int_{\mathbb{R}} x f(x)\, dx \tag{7}$$

i.e., the **integral** of the product of the possible value x taken by X and $f(x)$, the likelihood of X taking value x. In the homework, we will be asked to prove the **linearity of expectation** by using those definitions.

Another important concept is the **variance**, defined as

$$Var(X) = \mathbb{E}(X - \mathbb{E}X)^2 \tag{8}$$

the connection between variance and expectation can be given by the useful formula that

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 \tag{9}$$

for two random variables, we can define the **covariance** to describe their relationship

$$cov(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)] \tag{10}$$

and a similar identity holds that

$$cov(X) = \mathbb{E}XY - (\mathbb{E}X)(\mathbb{E}Y) \tag{11}$$

note that $cov(X, X) = Var(X)$ and that $cov(X, Y)$ is **bilinear**, i.e. $cov(aX + bY, Z) = a \cdot cov(X, Z) + b \cdot cov(Y, Z), cov(Z, aX + bY) = a \cdot cov(Z, X) + b \cdot cov(Z, Y)$ and **symmetric**, i.e. $cov(X, Y) = cov(Y, X)$. This is especially useful when computing the variance of a linear combination. For example, if we want to write $Var(2X + 3Y)$ in terms of $Var(X), Var(Y)$,

$$\begin{aligned}
Var(2X + 3Y) &= cov(2X + 3Y, 2X + 3Y) && (12)\\
&= 2cov(X, 2X + 3Y) + 3cov(Y, 2X + 3Y) && (13)\\
&= 2[2cov(X, X) + 3cov(X, Y)] + 3[2cov(Y, X) + 3cov(Y, Y)] && (14)\\
&= 4Var(X) + 12cov(X, Y) + 9cov(Y, Y) && (15)
\end{aligned}$$

you are asked to prove a more general version of this property in the homework.

Finally, let's talk about **normal distribution**. We say $X \sim N(\mu, \sigma^2)$ if it has density

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \ (x \in \mathbb{R}) \tag{16}$$

for the two parameters μ, σ^2 of normal distribution, a direct interpretation is that $\mathbb{E}X = \mu, Var(X) = \sigma^2$. You can try to prove those properties on your own by applying the definitions of expectation and variance to calculate the integrals. A trick will be that when calculating the integral

$$\mathbb{E}X = \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx \tag{17}$$

use the change of variables $u = \frac{x-\mu}{\sigma}$ to make the life easier

$$\begin{aligned}
\mathbb{E}X &= \int_{\mathbb{R}} x \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \, dx & (18)\\
&= \frac{1}{\sqrt{2\pi}\sigma} \sigma \int_{\mathbb{R}} (\sigma u + \mu) e^{-\frac{u^2}{2}} \, du & (19)\\
&= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} (\sigma u + \mu) e^{-\frac{u^2}{2}} \, du & (20)\\
&= \frac{\mu}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{u^2}{2}} \, du & (21)\\
&= \mu & (22)
\end{aligned}$$

here we use the property that $ue^{-\frac{u^2}{2}}$ is an odd function and that $\int_{\mathbb{R}} e^{-\frac{u^2}{2}} \, du = \sqrt{2\pi}$ (this property can be deduced from the standard normal density, en easy way to remember). The calculation of variance is left to the reader.

The **standard normal CDF** is one of the most frequently used notations in statistics. The definition is

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt = \mathbb{P}(G \le x) \ \ (G \sim N(0,1)) \tag{23}$$

a property of Φ is that

$$\forall x \in \mathbb{R}, \Phi(x) + \Phi(-x) = 1 \tag{24}$$

to see this, notice that $\frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$ is an even function in t, so

$$\begin{aligned}
\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} \, dt &= \int_{-x}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du \ (u = -t) & (25)\\
&= \int_{-\infty}^{+\infty} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du - \int_{-\infty}^{-x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} \, du & (26)\\
&= 1 - \Phi(-x) & (27)
\end{aligned}$$

Week 2

HW 0

For the problems in HW 0, let's look at problem 6 and 7 briefly. The important fact used in problem 6 is that for independent r.v. X, Y, it's true that $\mathbb{E}XY = \mathbb{E}X \cdot \mathbb{E}Y$. Let's prove this property for continuous random variables.

If X, Y are independent with density f, g, the joint density is $h(x, y) = f(x)g(y)$

$$\begin{aligned}
\mathbb{E}XY &= \int_{\mathbb{R}^2} xyh(x, y)\, dx\, dy && (28)\\
&= \int_{\mathbb{R}^2} xyf(x)g(y)\, dx\, dy && (29)\\
&= \int_{\mathbb{R}} xf(x)\, dx \cdot \int_{\mathbb{R}} yg(y)\, dy && (30)\\
&= \mathbb{E}X \cdot \mathbb{E}Y && (31)
\end{aligned}$$

one can also try to prove the property in the discrete case.

For problem 7, the main idea is to tell you that often it's the case that you can greatly simplify the calculations by applying the properties of expectation or variance. For $X \sim N(\mu_x, \sigma_x^2), Y \sim N(\mu_y, \sigma_y^2)$ independent, by independence, the joint density is

$$f(x, y) = f_X(x)f_Y(y) = \frac{1}{\sqrt{2\pi}\sigma_x} e^{-\frac{(x-\mu_x)^2}{2\sigma_x^2}} \frac{1}{\sqrt{2\pi}\sigma_y} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}} \tag{32}$$

the expectation can be calculated by linearity

$$\mathbb{E}(aX + bY + c) = a\mathbb{E}X + b\mathbb{E}Y + c = a\mu_x + b\mu_y + c \tag{33}$$

the second moment computed with the variance identity

$$\mathbb{E}X^2 = Var(X) + (\mathbb{E}X)^2 = \sigma_x^2 + \mu_x^2 \tag{34}$$

and the variance of linear combination is

$$Var(aX + bY + c) = Var(aX) + Var(bY) = a^2\sigma_x^2 + b^2\sigma_y^2 \tag{35}$$

note that generally the variance of sum does not equal the sum of variance, here it holds because of independence (actually this property holds if and only if X, Y are uncorrelated by the conclusion of problem 5).

HW 1

Let's talk about calculating the distribution of the transformation of a random variable. The most important idea comes from **the CDF method** that focuses on deriving the CDF of the transformed r.v.

To see how this method works, let's first look at some examples and then build up the theory for this method.

Now $X \sim N(0,1)$, and we want to derive the PDF of $Y = |X|$ and to calculate $\mathbb{E}|X|$. The first step is to set up the CDF of Y, denoted $F_Y(y) = \mathbb{P}(Y \leq y)$, it's obvious that when $y < 0$ the CDF always has value 0 so we only have to consider the non-trivial case where $y \geq 0$. Denote $f_X(x)$ as the density of X so $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

$$F_Y(y) = \mathbb{P}(|X| \leq y) = \mathbb{P}(-y \leq X \leq y) = \int_{-y}^{y} f_X(x)\, dx = 2\int_0^y f_X(x)\, dx \tag{36}$$

since $f_X(x)$ is an even function. Actually one does not have to calculate this integral, but to notice that PDF is the derivative of CDF, so taking derivative w.r.t. y on both sides gives

$$f_Y(y) = \frac{d}{dy} F_Y(y) = 2\frac{d}{dy}\int_0^y f_X(x)\, dx = 2f_X(y) = \sqrt{\frac{2}{\pi}} e^{-\frac{y^2}{2}} \ (y \geq 0) \tag{37}$$

the calculation of expectation follows

$$\begin{aligned}
\mathbb{E}Y &= \int_0^\infty y f_Y(y)\, dy && (38)\\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty y e^{-\frac{y^2}{2}}\, dy && (39)\\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-\frac{y^2}{2}}\, d\frac{y^2}{2} && (40)\\
&= \sqrt{\frac{2}{\pi}} \int_0^\infty e^{-u}\, du && (41)\\
&= \sqrt{\frac{2}{\pi}} && (42)
\end{aligned}$$

Remark. *Do not forget that the density for Y only works on $[0,\infty)$ so it's necessary to label out $y \geq 0$.*

Remark. *One might have to take the derivative of an integral with variables in the integration region a lot when calculating the distribution of the transformation of r.v. As a result, one might find the following property from calculus useful:*

$$\begin{aligned}
\frac{d}{dx}\int_{f(x)}^{g(x)} h(t)\, dt &= \frac{d}{dx}\int_0^{g(x)} h(t)\, dt - \frac{d}{dx}\int_0^{f(x)} h(t)\, dt && (43)\\
&= h(g(x))g'(x) - h(f(x))f'(x) && (44)
\end{aligned}$$

to see this, one can consider $p(x) = \int_0^x h(t)\,dt$ *and* $\int_0^{g(x)} h(t)\,dt = p(g(x))$*, so*

$$\frac{d}{dx}\int_0^{g(x)} h(t)\,dt = \frac{d}{dx}p(g(x)) \tag{45}$$

$$= p'(g(x))g'(x) \tag{46}$$

$$= h(g(x))g'(x) \tag{47}$$

since $p'(x) = h(x)$ *by Newton-Lebniz formula. So this is actually just an application of* ***the chain rule****.*

The example above tells us the way to apply the CDF method, now let's build up the method in theory. Let's assume that we already know the PDF of X and want to get the PDF of $Y = h(X)$ with h to be strictly monotone increasing (this assumption is made to simplify the proof but not necessary).

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(h(X) \leq y) \tag{48}$$

$$= \mathbb{P}\left(X \leq h^{-1}(y)\right) \tag{49}$$

$$= \int_{-\infty}^{h^{-1}(y)} f_X(x)\,dx \tag{50}$$

take derivative w.r.t. y on both sides to get

$$f_Y(y) = \frac{d}{dy}F_Y(y) \tag{51}$$

$$= \frac{d}{dy}\int_{-\infty}^{h^{-1}(y)} f_X(x)\,dx \tag{52}$$

$$= f_X(h^{-1}(y)) \cdot \frac{d}{dy}h^{-1}(y) \tag{53}$$

by the calculations we have already made in the remark above. (This is part of the homework problem 8, please try to prove the other half when h is strictly decreasing on your own) Notice that the density has to be non-negative and here since h is increasing, $\frac{d}{dy}h^{-1}(y)$ has to be non-negative, making the density f_Y non-negative. For the case where h is decreasing, there is a slight difference in the sign that you have to notice. In all, the general formula is given by

$$f_Y(y) = f_X(h^{-1}(y)) \cdot \left|\frac{d}{dy}h^{-1}(y)\right| \tag{54}$$

for **any strictly monotone** h and is called **the transformation method**.

Remark. *Although this method directly comes from the CDF method, one will see that in multi-dimensional case this method is much easier to generalize and to apply.*

Now let me raise an example to show you how to apply this method. Consider $X \sim N(\mu, \sigma^2)$ and we want to find the PDF of $Y = \frac{X-\mu}{\sigma}$. It's immediate that $h(x) = \frac{x-\mu}{\sigma}$ is a linear function so it's strictly monotone, $h^{-1}(y) = \sigma y + \mu$

and $\frac{d}{dy}h^{-1}(y) = \frac{1}{\frac{dh(y)}{dy}} = \sigma$. Now it's clear that $f_X(x) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, plug in the formula to see

$$
\begin{aligned}
f_Y(y) &= f_X(h^{-1}(y)) \cdot \left| \frac{d}{dy}h^{-1}(y) \right| && (55)\\
&= \sigma \cdot f_X(\sigma y + \mu) && (56)\\
&= \frac{1}{\sqrt{2\pi}}e^{-\frac{y^2}{2}} && (57)
\end{aligned}
$$

as a result, we see that $Y \sim N(0,1)$ is standard Gaussian. So we proved a very important **scaling property of Gaussian random variable** that $X \sim N(\mu, \sigma^2)$ if and only if $\frac{X-\mu}{\sigma} \sim N(0,1)$.

The last method to talk about is **the method of moment generating function (MGF)**. This method is based on two properties of MGF defined as $M_X(t) = \mathbb{E}e^{tX}$. The first one is that MGF characterizes the distribution, so two random variables have the same MGF if and only if they have the same distribution. The second one is that for independent X, Y, $M_{X+Y}(t) = M_X(t)M_Y(t)$ the MGF of the sum is the product of respective MGF. **The MGF method is especially effective for dealing with Gaussian random variables.**

An example is that for $Y_1, Y_2, ..., Y_n \sim N(0,1)$ *i.i.d.*, let's calculate the distribution of $Z = a_1Y_1 + a_2Y_2 + ... + a_nY_n$. One has to know that the MGF for $N(\mu, \sigma^2)$ Gaussian r.v. is $M(t) = e^{\mu t + \frac{\sigma^2}{2}t^2}$ (refer to the remark is you are not familiar with this conclusion).

$$
\begin{aligned}
M_Z(t) &= M_{a_1Y_1}(t)M_{a_2Y_2}(t)...M_{a_nY_n}(t) && (58)\\
&= \mathbb{E}e^{ta_1Y_1}\mathbb{E}e^{ta_2Y_2}...\mathbb{E}e^{ta_nY_n} && (59)\\
&= M_{Y_1}(ta_1)...M_{Y_n}(ta_n) && (60)\\
&= e^{\frac{a_1^2}{2}t^2}...e^{\frac{a_n^2}{2}t^2} && (61)\\
&= e^{\frac{\sum_{i=1}^n a_i^2}{2}t^2} && (62)
\end{aligned}
$$

comparing with the MGF for $N(\mu, \sigma^2)$, one immediately find that $Z \sim N(0, \sum_{i=1}^n a_i^2)$. This is telling us that **the linear combination of independent Gaussian r.v. must still be Gaussian**. (Try to do the same problem for $Y_i \sim N(\mu_i, \sigma_i^2)$ independent but not *i.i.d.* to see the conclusion that the linear combination is still Gaussian)

Remark. *Let's calculate the MGF for $X \sim N(\mu, \sigma^2)$*

$$
\begin{aligned}
M_X(t) &= \mathbb{E}e^{tX} && (63)\\
&= \int_{\mathbb{R}} e^{tx} f_X(x)\,dx && (64)\\
&= \frac{1}{\sqrt{2\pi}\sigma}\int_{\mathbb{R}} e^{tx - \frac{(x-\mu)^2}{2\sigma^2}}\,dx \quad \left(u = \frac{x-\mu}{\sigma}\right) && (65)\\
&= \frac{1}{\sqrt{2\pi}}\int_{\mathbb{R}} e^{t(\sigma u + \mu) - \frac{u^2}{2}}\,du && (66)
\end{aligned}
$$

extract the constant term $e^{t\mu}$ *to continue*

$$\begin{aligned}
M_X(t) &= \frac{1}{\sqrt{2\pi}} e^{t\mu} \int_{\mathbb{R}} e^{t\sigma u - \frac{u^2}{2}}\, du && (67)\\
&= \frac{1}{\sqrt{2\pi}} e^{t\mu} \int_{\mathbb{R}} e^{-\frac{1}{2}(u^2 - 2t\sigma u + t^2\sigma^2) + \frac{\sigma^2}{2}t^2}\, du && (68)\\
&= \frac{1}{\sqrt{2\pi}} e^{\mu t + \frac{\sigma^2}{2}t^2} \int_{\mathbb{R}} e^{-\frac{(u-t\sigma)^2}{2}}\, du \ (v = u - t\sigma) && (69)\\
&= \frac{1}{\sqrt{2\pi}} e^{\mu t + \frac{\sigma^2}{2}t^2} \int_{\mathbb{R}} e^{-\frac{v^2}{2}}\, dv && (70)\\
&= e^{\mu t + \frac{\sigma^2}{2}t^2} && (71)
\end{aligned}$$

Week3

It's always important that when calculating the distribution of the transformation of random variables, one choose the method that fits the best with the problem. Now we have three methods: the CDF method, the transformation method and the MGF method.

Generally, when dealing with the distribution of the sum of independent random variables, always use MGF method. The reason is that the MGF of the sum of independent r.v. is always the product of respective MGF. From the one-to-one correspondence between MGF and distribution, one would always find out the distribution of the independent sum easily.

For transformation method, it's always easy to apply when we see a transformation from $\mathbb{R}^n \to \mathbb{R}^n$, mapping a random vector of length n to another random vector of length n. The key point is that the dimension of the domain and image space of the transformation should be the same (because it depends on the determinant of the Jacobian as we will see later). As a result, transformation method won't be applied for problem like deriving the distribution of the sum of random variables, since it's actually mapping $(X_1, ..., X_n) \in \mathbb{R}^n$ to $X_1 + ... + X_n \in \mathbb{R}$. Moreover, there's some restrictions on the 'invertible' property of the transformation. For example, $Y = X^2$ has transformation $h(x) = x^2$ which is not invertible, so the transformation method will fail.

For distribution method, it's the most general method but also the method with the most calculations involved. It can be applied in all circumstances, regardless of the transformation function and the random variables one is using. Problems like $Y = |X|, Y = X^2$ can only be dealt with using the CDF method.

Let's look at some problems that consider the distribution of the average of *i.i.d.* random variables $\frac{S_n}{n} = \frac{X_1+...+X_n}{n}$ to get familiar with the MGF method.

Let's first take X_1 following the Bernoulli distribution $B(1, p)$. Let's first calculate the MGF of X_1

$$M_{X_1}(t) = \mathbb{E}e^{tX_1} = 1 - p + pe^t \tag{72}$$

so now we see that

$$M_{\frac{S_n}{n}}(t) = \mathbb{E}e^{\frac{S_n}{n}t} = \mathbb{E}e^{\frac{t}{n}S_n} = M_{S_n}\left(\frac{t}{n}\right) \tag{73}$$

since S_n is *i.i.d.* sum, its MGF is the product of the respective MGF, so

$$M_{S_n}\left(\frac{t}{n}\right) = \left[M_{X_1}\left(\frac{t}{n}\right)\right]^n = (1 - p + pe^{\frac{t}{n}})^n \tag{74}$$

notice the trick to put the denominator n of $\frac{S_n}{n}$ into the variable in the MGF.

Similarly, we can compute the example for Poisson distribution $X_1 \sim P(\lambda)$

$$M_{X_1}(t) = \mathbb{E}e^{tX_1} = \sum_{k=0}^{\infty} e^{tk}\frac{\lambda^k}{k!}e^{-\lambda} = e^{\lambda e^t - \lambda} \tag{75}$$

so now we see that

$$M_{\frac{S_n}{n}}(t) = \mathbb{E}e^{\frac{S_n}{n}t} = \mathbb{E}e^{\frac{t}{n}S_n} = M_{S_n}\left(\frac{t}{n}\right) \tag{76}$$

since S_n is *i.i.d.* sum, its MGF is the product of the respective MGF, so

$$M_{S_n}\left(\frac{t}{n}\right) = \left[M_{X_1}\left(\frac{t}{n}\right)\right]^n = e^{n\lambda e^{\frac{t}{n}} - n\lambda} \tag{77}$$

one might try to prove the additivity of Poisson distribution as an exercise (if $\forall i = 1, 2, ..., n, X_i \sim P(\lambda_i)$ are independent random variables, then $X_1 + ... + X_n \sim P(\lambda_1 + ... + \lambda_n)$) using MGF.

Week 4

Quiz Answer

1. We are playing a game of darts, where every throw results in the dart landing randomly somewhere on the dartboard, which has a radius of 1. If we say that X is the horizontal coordinate and Y is the vertical coordinate (both measured from the center/bullseye), then each throw results in a random pair (X, Y) with a joint density of

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{\pi} & x^2+y^2 \le 1 \\ 0 & else \end{cases} \tag{78}$$

(a) (6 points) What is the probability we are closer to the bullseye in the horizontal direction than in the vertical direction?

solution. The probability is $\mathbb{P}(|X| < |Y|)$. Representing as the integral of joint density

$$\begin{aligned} \mathbb{P}(|X| < |Y|) &= \frac{1}{\pi} \int\int_{x^2+y^2\le 1, |x|<|y|} dx\, dy \qquad (79) \\ &= \frac{1}{\pi} \cdot area(x^2+y^2 \le 1, |x| < |y|) \qquad (80) \\ &= \frac{1}{\pi} \cdot \frac{\pi}{2} = \frac{1}{2} \qquad (81) \end{aligned}$$

□

(b) (4 points) What is the marginal density of the vertical component, $f_Y(y)$

solution.

$$\begin{aligned} f_Y(y) &= \int_{-\sqrt{1-y^2}}^{\sqrt{1-y^2}} f_{X,Y}(x,y)\, dx \qquad (82) \\ &= \frac{2}{\pi}\sqrt{1-y^2},\ y \in (-1,1) \qquad (83) \end{aligned}$$

□

2. Suppose that at the start of the week, the amount of gasoline (measured in 1K gallons) in the holding tank of our neighborhood gas station can be represented by a random variable, G, with a uniform distribution over [0, 1] (so the maximum possible is 1000 gallons of gasoline at the start of the week). Further, suppose that the amount of gasoline sold by the gas station during the week can also be modeled with uniform random variable, Y , over the interval $[0, g_1]$, where g_1 is the particular amount at the start of the week.

(a) (5 points) Find the joint density function for the amount at the start of the week and the amount sold during the week.

solution. Now we know that $G \sim U(0,1), Y|_{G=g_1} \sim U(0, g_1)$ so

$$f_{Y,G}(y,g) = f_G(g) \cdot f_{Y|G}(y|g) = \frac{1}{g},\ 0 < y < g < 1 \tag{84}$$

□

(b) (5 points) If the station stocks 600 gallons at the start of the week, what is the probability they sell more than 250 gallons?

solution. Condition on $G = 0.6$, $Y \sim U(0, 0.6)$, so

$$\mathbb{P}\left(Y \geq 0.25 | G = 0.6\right) = \int_{0.25}^{0.6} \frac{1}{0.6}\, dy = \frac{0.35}{0.6} = \frac{7}{12} \tag{85}$$

□

(c) (5 points) Find the marginal density of Y.

solution.

$$f_Y(y) = \int_y^1 f_{Y,G}(y,g)\, dg = \int_y^1 \frac{1}{g}\, dg = -\log y,\ y \in (0,1) \tag{86}$$

□

(d) (5 points) If we know that the station sold 250 gallons of gasoline, what is the probability that they had more than 500 gallons at the start of the week?

solution. In order to get $\mathbb{P}\left(G \geq 0.5 | Y = 0.25\right)$, let's first derive the conditional density $f_{G|Y}(g|y)$

$$f_{G|Y}(g|y) = \frac{f_{Y,G}(y,g)}{f_Y(y)} = \frac{1}{-\log y \cdot g}, 0 < y < g < 1 \tag{87}$$

so now

$$\mathbb{P}\left(G \geq 0.5 | Y = 0.25\right) = \int_{0.5}^1 f_{G|Y}(g|0.25)\, dg \tag{88}$$

$$= \int_{0.5}^1 \frac{1}{-\log 0.25 \cdot g}\, dg \tag{89}$$

$$= \frac{\log 0.5}{\log 0.25} = \frac{1}{2} \tag{90}$$

□

(e) (5 points) What is the expected amount of gasoline we will have left at the end of the week?

solution.

$$\mathbb{E}(G-Y) = \mathbb{E}G - \mathbb{E}Y \tag{91}$$

where $\mathbb{E}G = \frac{1}{2}$ and $\mathbb{E}Y = \mathbb{E}[\mathbb{E}(Y|G)] = \mathbb{E}\frac{G}{2} = \frac{1}{4}$, so $\mathbb{E}(G-Y) = \frac{1}{4}$. □

3. The hens at Lilly's Cage-free Egg Farm produce eggs according to a Poisson distribution, with a mean of β per day. Because the hens are allowed to roam free, and not kept in cages, they lay their eggs around the barnyard which increases the chance that they will break. The eggs survive to be collected with some probability p.

(a) (5 points) Find the expected number of eggs that will survive to be collected.

solution. Assume there are $N \sim P(\beta)$ eggs laid and C eggs collected. So $C|_{N=n} \sim B(n,p)$.

$$\mathbb{E}C = \mathbb{E}[\mathbb{E}(C|N)] = \mathbb{E}pN = p\mathbb{E}N = p\beta \tag{92}$$

□

(b) (5 points) Find the variance for the number of eggs that will survive to be collected.

solution.

$$\mathbb{E}C^2 = \mathbb{E}[\mathbb{E}(C^2|N)] \tag{93}$$

now $\mathbb{E}(C^2|N) = Var(C|N) + [\mathbb{E}(C|N)]^2 = Np(1-p) + N^2p^2$ so

$$\begin{aligned}
\mathbb{E}C^2 = \mathbb{E}[\mathbb{E}(C^2|N)] &= \mathbb{E}Np(1-p) + \mathbb{E}N^2p^2 && (94)\\
&= p(1-p)\mathbb{E}N + p^2\mathbb{E}N^2 && (95)\\
&= p(1-p)\beta + p^2(\beta^2+\beta) && (96)
\end{aligned}$$

since $\mathbb{E}N^2 = Var(N) + (\mathbb{E}N)^2 = \beta^2 + \beta$, so

$$\begin{aligned}
Var(C) = \mathbb{E}C^2 - (\mathbb{E}C)^2 &= p(1-p)\beta + p^2(\beta^2+\beta) - p^2\beta^2 && (97)\\
&= p(1-p)\beta + p^2\beta = p\beta && (98)
\end{aligned}$$

□

Central Limit Theorem

CLT only works for *i.i.d.* **random variable series that has finite second moment**. The limiting distribution of $\frac{S_n - \mathbb{E}S_n}{\sqrt{Var(S_n)}}$ is always $N(0,1)$. When one wants to apply CLT, first verify that the random variable series is *i.i.d.*, then verify that second moment exists. If those two conditions hold, one only need to calculate the expectation and the variance of the sum $S_n = X_1 + ... + X_n$ to write out the normal approximation.

For example, if we have $X_1, ..., X_n \sim P(\lambda)$ *i.i.d.*, since $Var(X_1) = \lambda < \infty$, CLT holds and we compute

$$\mathbb{E}S_n = n\mathbb{E}X_1 = n\lambda \tag{99}$$

$$Var(S_n) = nVar(X_1) = n\lambda \tag{100}$$

to conclude that

$$\frac{S_n - n\lambda}{\sqrt{n\lambda}} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{101}$$

which means that when n is large enough, the following approximation that

$$\mathbb{P}\left(\frac{S_n - n\lambda}{\sqrt{n\lambda}} \leq x\right) \to \Phi(x) \ (n \to \infty) \tag{102}$$

works for standard Gaussian CDF Φ. If we simplify the expression, we can see that $\sqrt{n}(\overline{X}_n - \lambda) \xrightarrow{d} N(0, \lambda)$.

Now if we have paired observations $X_1, ..., X_n, Y_1, ..., Y_n$ to be independent and X_i all have the same distribution $N(\mu_1, \sigma_1^2)$, Y_i all have the same distribution $N(\mu_2, \sigma_2^2)$, then we can consider $S_n = (X_1 - Y_1) + ... + (X_n - Y_n)$ to see that $X_1 - Y_1, ..., X_n - Y_n$ are *i.i.d.* with $Var(X_1 - Y_1) = \sigma_1^2 + \sigma_2^2 < \infty$ so CLT holds. Compute

$$\mathbb{E}S_n = n(\mu_1 - \mu_2) \tag{103}$$

$$Var(S_n) = n(\sigma_1^2 + \sigma_2^2) \tag{104}$$

to get the conclusion that

$$\frac{S_n - n(\mu_1 - \mu_2)}{\sqrt{n(\sigma_1^2 + \sigma_2^2)}} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{105}$$

divide numerator and denominator by n to see

$$\frac{\sqrt{n}[(\overline{X}_n - \overline{Y}_n) - (\mu_1 - \mu_2)]}{\sqrt{(\sigma_1^2 + \sigma_2^2)}} \xrightarrow{d} N(0,1) \ (n \to \infty) \tag{106}$$

Remark. *The simplest CLT does not hold for $X_1, ..., X_n, Y_1, ..., Y_n$ since they are independent but not identically distributed. However, if we notice the fact that they are paired samples, we can consider the difference $X_i - Y_i$ as a new random variable series so now it's i.i.d..*

Week 5

Bias Variance Decomposition

For any estimator $\hat{\theta} = \hat{\theta}(X), X = (X_1, ..., X_n)$ that estimates the true parameter θ, we hope to set up a criterion for selecting the best estimator. A frequently used criterion is the mean square error

$$MSE(\hat{\theta}) = \mathbb{E}(\hat{\theta} - \theta)^2 \tag{107}$$

and it's important to realize that the mean square error always has the bias variance decomposition

$$\begin{aligned} MSE(\hat{\theta}) &= \mathbb{E}(\hat{\theta} - \theta)^2 && (108) \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta} + \mathbb{E}\hat{\theta} - \theta)^2 && (109) \\ &= \mathbb{E}(\hat{\theta} - \mathbb{E}\hat{\theta})^2 + \mathbb{E}(\mathbb{E}\hat{\theta} - \theta)^2 + 2\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)] && (110) \\ &= Var(\hat{\theta}) + Bias^2(\hat{\theta}) && (111) \end{aligned}$$

since $\mathbb{E}[(\hat{\theta} - \mathbb{E}\hat{\theta})(\mathbb{E}\hat{\theta} - \theta)] = (\mathbb{E}\hat{\theta} - \theta)(\mathbb{E}\hat{\theta} - \mathbb{E}\hat{\theta}) = 0$. This is showing some kind of **trade-off between unbiasedness and efficiency under a fixed MSE**. If one wants to find an unbiased estimator, one always has to sacrifice some kind of efficiency, while if one wants to find an efficient estimator, it may be seriously biased(the trivial estimator $\hat{\theta} = 0$ always has zero variance but may be significantly biased).

Let's see an example in the textbook exercise 8.20 where $Y_1, ..., Y_4 \sim \mathcal{E}\left(\frac{1}{\theta}\right)$ and we want to estimate the parameter θ. Now $X = \sqrt{Y_1 Y_2}$ and we want to find a multiple of X which is unbiased.

Let's first calculate $\mathbb{E}X = \mathbb{E}^2\sqrt{Y_1}$. Note that

$$\begin{aligned} \mathbb{E}\sqrt{Y_1} &= \int_0^\infty \sqrt{y}\frac{1}{\theta}e^{-\frac{y}{\theta}}\, dy && (112) \\ &= \sqrt{\theta}\int_0^\infty \sqrt{u}e^{-u}\, du \ \left(u = \frac{y}{\theta}\right) && (113) \\ &= \sqrt{\theta} \cdot \Gamma\left(\frac{3}{2}\right) && (114) \\ &= \sqrt{\theta} \cdot \frac{1}{2}\Gamma\left(\frac{1}{2}\right) && (115) \\ &= \frac{\sqrt{\theta\pi}}{2} && (116) \end{aligned}$$

as a result, $\mathbb{E}X = \frac{\theta\pi}{4}$ and $\mathbb{E}\frac{4}{\pi}\sqrt{Y_1 Y_2} = \theta$ so $\frac{4}{\pi}\sqrt{Y_1 Y_2}$ is unbiased.

Remark. *Recall from calculus the form of Gamma function that*

$$\Gamma(s) = \int_0^\infty x^{s-1}e^{-x}\, dx \ (s > 0) \tag{117}$$

and the property that $\forall p \in (0,1), \Gamma(p)\Gamma(1-p) = \frac{\pi}{\sin p\pi}$ *so* $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$ *(or directly from the hint).*

For another example, refer to exercise 8.36 in the textbook. $Y_1, ..., Y_n$ follow $\mathcal{E}\left(\frac{1}{\theta}\right)$ so we now that

$$\mathbb{E}Y_1 = \theta, Var(Y_1) = \theta^2, \mathbb{E}\overline{Y} = \theta, Var(\overline{Y}) = \frac{\theta^2}{n} \tag{118}$$

now we want to construct unbiased estimator for θ and provide an estimate for the standard error of this estimator.

Now we want the expectation of the estimator to be θ so it's quite obvious that Y_1 is just an unbiased estimator. To find the standard error

$$se(Y_1) = \theta \tag{119}$$

since θ it self is unknown, it can be estimated in various ways. We can take $\overline{Y}$ as an unbiased estimation of θ so the estimated standard error is

$$\hat{se}(Y_1) = \overline{Y} \tag{120}$$

we can also take sample variance S^2 as an estimation of θ^2 so the estimated standard error is

$$\hat{se}(Y_1) = \sqrt{S^2} \tag{121}$$

there's no fixed way to estimate the standard error so you can put up any reasonable ways to do it! On the other hand, the choice of unbiased estimator is also not unique, Y_1 is unbiased, but $\overline{Y}$ is also unbiased, one may also choose the sample mean as the operator and try to calculate the estimated standard error.

Consistency

Estimator $\hat{\theta}$ is called a consistent estimator if

$$\hat{\theta} \xrightarrow{p} \theta \ (n \to \infty) \tag{122}$$

which means that $\forall \varepsilon > 0, \mathbb{P}\left(|\hat{\theta} - \theta| \geq \varepsilon\right) \to 0 \ (n \to \infty)$, this convergence is called convergence in probability.

Since convergence in probability is preserved under addition, subtraction, multiplication, division and actually any continuous mapping (continuous mapping theorem), consistency is typically easy to get. In other words, if $\hat{\theta}_1, \hat{\theta}_2$ are consistent estimators of θ, then $\frac{\hat{\theta}_1 + \hat{\theta}_2}{2}$ is still a consistent estimator.

By what we have learnt so far, sample variance S^2 is an unbiased consistent estimator of population variance σ^2, so S **is a consistent estimator of** σ (apply the continuous function $g(x) = \sqrt{x}$ on both sides). However, note that S **is not an unbiased estimator of** σ! (since generally $\mathbb{E}\sqrt{X} \neq \sqrt{\mathbb{E}X}$)

We raise and example next to illustrate the way to construct a consistent estimator and to prove its consistency. For sample $Y_1, ..., Y_n$ with $\mathbb{E}Y_1 = \mu, Var(Y_1) = \sigma^2$, we want to estimate $\mathbb{E}Y_1^2$ consistently.

A natural estimator comes from the sample mean of second moments

$$T = \frac{\sum_{i=1}^{n} Y_i^2}{n} \tag{123}$$

and to show its consistency, it directly follows from WLLN that since $\mathbb{E}Y_1^2 < \infty$

$$T \xrightarrow{p} \mathbb{E}Y_1^2 \ (n \to \infty) \tag{124}$$

Efficiency

We say an estimator is more efficient than the other estimator is it has lower variance with the relative efficiency defined as

$$\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{Var(\hat{\theta}_2)}{Var(\hat{\theta}_1)} \tag{125}$$

in this course we just need to learn to calculate the relative efficiency between two given estimators.

For example, if $Y_1, ..., Y_n$ are from a population with mean μ and variance σ^2, if we have $\hat{\mu}_2 = \frac{1}{4}Y_1 + \frac{Y_2+...+Y_{n-1}}{2(n-2)} + \frac{1}{4}Y_n$ and $\hat{\mu}_3 = \overline{Y}$, then

$$Var(\hat{\mu}_3) = \frac{Var(Y_1)}{n} = \frac{\sigma^2}{n} \tag{126}$$

$$Var(\hat{\mu}_2) = \frac{1}{16}\sigma^2 + \frac{n-2}{4(n-2)^2}\sigma^2 + \frac{1}{16}\sigma^2 = \frac{1}{8}\sigma^2 + \frac{1}{4(n-2)}\sigma^2 \tag{127}$$

so

$$\text{eff}(\hat{\mu}_3, \hat{\mu}_2) = \frac{Var(\hat{\mu}_2)}{Var(\hat{\mu}_3)} = \frac{n}{8} + \frac{n}{4(n-2)} \tag{128}$$