STAT 7032 Probability CLT part

Wlodek Bryc

Created: Friday, Jan 2, 2014Printed: April 24, 2020 `File: Grad-Prob-2020-slides.TEX`

April 24, 2020

Facts to use
$\varphi(t) = E\exp(itX)$

- For standard normal distribution $\varphi(t) = e^{-t^2/2}$
- The following are equivalent:
  - $X_n \xrightarrow{\mathcal{D}} X$
  - $\varphi_n(t) \to \varphi(t)$ for all $t \in \mathbb{R}$.
- If X is square integrable with mean zero and variance σ^2 then

$$\left|\varphi(t) - (1 - \tfrac{\sigma^2 t^2}{2})\right| \leq E(\min\{\frac{1}{6}|tX|^3, (tX)^2\}) \tag{1}$$

Proof: $\varphi(t) = Ee^{-itX}$. This relies on two integral identities applied to $x = tX(\omega)$ under the integral: $\left|e^{ix} - (1 + ix - \frac{x^2}{2})\right| = \left|\frac{i}{2}\int_0^x (x-s)^2 e^{is} ds\right| \leq \frac{|x^3|}{6}$ $\left|e^{ix} - (1 + ix - \frac{x^2}{2})\right| = \left|\int_0^x (x-s)(e^{is}-1) ds\right| \leq x^2$ $\square$

Last time we used inequality $|z_1^n - z_2^n| \leq n|z_1 - z_2|$ complex numbers of modulus at most 1 which we now generalize.

Lemma 1. *If $z_1, \dots, z_m$ and $w_1, \dots, w_m$ are complex numbers of modulus at most 1 then*

$$|z_1 \dots z_m - w_1 \dots w_m| \leq \sum_{k=1}^m |z_k - w_k| \tag{2}$$

Proof. Write the left hand side of (2) as a telescoping sum:

$$\begin{aligned} z_1 \dots z_m - w_1 \dots w_m &= z_1 \dots z_m - w_1 z_2 \dots z_m + w_1 z_2 \dots z_m - w_1 w_2 \dots z_m \\ &\quad \dots + w_1 w_2 \dots w_{m-1} z_m - w_1 w_2 \dots w_m \\ &= \sum_{k=1}^m w_1 \dots w_{k-1}(z_k - w_k) z_{k+1} \dots z_m \end{aligned}$$

$\square$

1 Lindeberg's theorem

Lindeberg's theorem

For each n we have a triangular array of random variables that are independent in each row

$$\begin{array}{l} X_{1,1}, X_{1,2}, \quad \dots \quad , X_{1,r_1} \\ X_{2,1}, X_{2,2}, \quad \dots \quad , X_{2,r_2} \\ \vdots \\ X_{n,1}, X_{n,2}, \quad \dots \quad , X_{n,r_n} \\ \vdots \end{array}$$

and we set $S_n = X_{n,1} + \cdots + X_{n,r_n}$. We assume that random variables are square-integrable with mean zero, and we use the notation

$$E(X_{n,k}) = 0,\ \sigma_{nk}^2 = E(X_{n,k}^2),\ s_n^2 = \sum_{k=1}^{r_n} \sigma_{nk}^2 \tag{3}$$

Definition 2 (The Lindeberg condition)**.** We say that the *Lindeberg condition* holds if

$$\forall_{\varepsilon>0} \lim_{n\to\infty} \frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP = 0 \tag{4}$$

Remark 3 (Important Observation)**.** *Under the Lindeberg condition, we have*

$$\lim_{n\to\infty} \max_{k\le r_n} \frac{\sigma_{nk}^2}{s_n^2} = 0 \tag{5}$$

Proof.

$$\sigma_{nk}^2 = \int_{|X_{nk}|\le\varepsilon s_n} X_{nk}^2 dP + \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP \le \varepsilon s_n^2 + \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP$$

So

$$\max_{k\le r_n} \frac{\sigma_{nk}^2}{s_n^2} \le \varepsilon + \frac{1}{s_n^2} \max_{k\le r_n} \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP$$

$$\le \varepsilon + \frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP$$

□

Theorem 4 (Lindeberg CLT)**.** *Suppose that for each n the sequence $X_{n1} \dots X_{n,r_n}$ is independent with mean zero. If the Lindeberg condition holds for all $\varepsilon > 0$ then $S_n/s_n \xrightarrow{\mathcal{D}} Z$.*

Example 5 (Suppose $X_1, X_2, \dots,$ are iid mean m variance $\sigma^2 > 0$. Then $S_n = \frac{1}{\sigma\sqrt{n}} \sum_{k=1}^n (X_k - m) \xrightarrow{\mathcal{D}} Z$.). •

Triangular array: $X_{n,k} = \frac{X_k - m}{\sqrt{n}\sigma}$ and $s_n = 1$.

- The Lindeberg condition is

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \int_{|X_k-m|>\varepsilon\sigma\sqrt{n}} \frac{(X_k-m)^2}{\sigma^2} dP$$

$$= \lim_{n\to\infty} \frac{1}{\sigma^2} \int_{|X_1-m|>\varepsilon\sigma\sqrt{n}} (X_1-m)^2 dP = 0$$

by Lebesgue dominated convergence theorem.

Proof of Lindeberg CLT I

Without loss of generality we may assume that $s_n^2 = 1$ so that $\sum_{k=1}^{r_n} \sigma_{nk}^2 = 1$.

- Denote $\varphi_{nk} = E(e^{itX_{nk}})$. By (1) we have

$$\left|\varphi_{nk}(t) - (1 - \frac{1}{2}t^2\sigma_{nk}^2)\right| \le E\left(\min\{|tX_{nk}|^2, |tX_{nk}|^3\}\right)$$

$$\le \int_{|X_{nk}|\le\varepsilon} |tX_{nk}|^3 dP + \int_{|X_{nk}|>\varepsilon} |tX_{nk}|^2 dP$$

$$\le t^3\varepsilon \int_{|X_{nk}|dP\le\varepsilon} X_{nk}^2 dP + t^2 \int_{|X_{nk}|>\varepsilon} X_{nk}^2 dP \le t^3\varepsilon\sigma_{nk}^2 + t^2 \int_{|X_{nk}|>\varepsilon} X_{nk}^2 dP \quad (6)$$

- Using (2), $\boxed{|z_1 \dots z_m - w_1 \dots w_m| \le \sum_{k=1}^m |z_k - w_k|}$ we see that for n large enough so that $\frac{1}{2}t^2\sigma_{nk}^2 < 1$

$$\left|\varphi_{S_n}(t) - \prod_{k=1}^{r_n}(1 - \tfrac{1}{2}t^2\sigma_{nk}^2)\right|$$

$$\le \varepsilon t^3 \sum_{k=1}^{r_n} \sigma_{nk}^2 + t^2 \sum_{k=1}^{r_n} \int_{|X_{nk}|>\varepsilon} X_{nk}^2 dP$$

Proof of Lindeberg CLT II

Since $\varepsilon > 0$ is arbitrary and $t \in \mathbb{R}$ is fixed, this shows that

$$\lim_{n\to\infty} \left|\varphi_{S_n}(t) - \prod_{k=1}^{r_n}(1 - \tfrac{1}{2}t^2\sigma_{nk}^2)\right| = 0$$

It remains to verify that $\lim_{n\to\infty} \left|e^{-t^2/2} - \prod_{k=1}^{r_n}(1 - \frac{1}{2}t^2\sigma_{nk}^2)\right| = 0$.

To do so, we apply the previous proof to the triangular array $Z_{n,k} = \sigma_{n,k} Z_k$ of independent normal random variables. Note that

$$\varphi_{\sum_{k=1}^{r_n} Z_{nk}}(t) = \prod_{k=1}^{r_n} e^{-t^2\sigma_{nk}^2/2} = e^{-t^2/2}$$

We only need to verify the Lindeberg condition for $\{Z_{nk}\}$.

Proof of Lindeberg CLT III

$$\int_{|Z_{nk}|>\varepsilon} Z_{nk}^2 dP = \sigma_{nk}^2 \int_{|x|>\varepsilon/\sigma_{nk}} x^2 f(x) dx$$

So for $\varepsilon > 0$ we estimate (recall that $\sum_k \sigma_{nk}^2 = 1$)

$$\sum_{k=1}^{r_n} \int_{|Z_{nk}|>\varepsilon} Z_{nk}^2 dP \le \sum_{k=1}^{r_n} \sigma_{nk}^2 \int_{|x|>\varepsilon/\sigma_{nk}} x^2 f(x) dx$$

$$\le \max_{1\le k\le r_n} \int_{|x|>\varepsilon/\sigma_{nk}} x^2 f(x) dx$$

$$= \int_{|x|>\varepsilon/\max_k \sigma_{nk}} x^2 f(x) dx$$

The right hand side goes to zero as $n \to \infty$, because by $\max_{1\le k\le r_n} \sigma_{nk} \to 0$ by (5). QED

2 Lyapunov's theorem

Lyapunov's theorem

Theorem 6. *Suppose that for each n the sequence $X_{n1} \dots X_{n,r_n}$ is independent with mean zero. If the* Lyapunov's condition

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{k=1}^{r_n} E|X_{nk}|^{2+\delta} = 0 \tag{7}$$

holds for some $\delta > 0$, then $S_n/s_n \xrightarrow{\mathcal{D}} Z$

Proof. We use the following bound to verify Lindeberg's condition:

$$\frac{1}{s_n^2} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\varepsilon s_n} X_{nk}^2 dP \le \frac{1}{\varepsilon^\delta s_n^{2+\delta}} \sum_{k=1}^{r_n} \int_{|X_{nk}|>\varepsilon s_n} |X_{nk}|^{2+\delta} dP$$

$$\le \frac{1}{\varepsilon^\delta s_n^{2+\delta}} \sum_{k=1}^{r_n} E|X_{nk}|^{2+\delta}$$

□

Corollary 7. *Suppose X_k are independent with mean zero, variance σ^2 and that $\sup_k E|X_k|^{2+\delta} < \infty$. Then $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} \sigma Z$.*

Proof. Let $C = \sup_k E|X_k|^{2+\delta}$. WLOG $\sigma > 0$. Then $s_n = \sigma\sqrt{n}$ and $\frac{1}{s_n^{2+\delta}} \sum_{k=1}^n E(|X_k|^{2+\delta}) \le \frac{Cn}{\sigma^{2+\delta} n^{1+\delta/2}} = \frac{C}{\sigma^{2+\delta} n^{\delta/2}} \to 0$, so Lyapunov's condition is satisfied. □

Corollary 8. *Suppose X_k are independent, uniformly bounded, and have mean zero. If $\sum_n Var(X_n) = \infty$, then $S_n/\sqrt{Var(S_n)} \xrightarrow{\mathcal{D}} N(0,1)$.*

Proof. Suppose $|X_n| \le C$ for a constant C. Then

$$\frac{1}{s_n^3} \sum_{k=1}^n E|X_n|^3 \le C \frac{s_n^2}{s_n^3} = \frac{C}{s_n} \to 0$$

□

The end
Lets stop here

- ~~Homework 11, due Monday - two exercises from Ch 11 of the notes.~~
- There is also a sizeable list of exercises from past prelims
- Things to do on Friday:
  - CLT without Lindeberg condition, when normalization is not by variance
  - Multivariate characteristic functions and multivariate normal distribution.

Thank you

Normal approximation without Lindeberg condition

3 Normal approximation without Lindeberg condition

Asymptotic normality may hold without Lindeberg condition:

- Normalization might be different that the variance. *In fact, the variance might be infinite!*

  A basic remedy for issues with the variance is Slutsky's theorem.

  - Truncation makes variances finite: $X_k = X_k I_{|X_k|\le a_n} + X_k I_{|X_k|>a_n}$
  - We use CLT for truncated r.v. $\frac{1}{s_n}\sum_{k=1}^n X_k I_{|X_k|\le a_n} \xrightarrow{\mathcal{D}} N(0,1)$ (triangular array)
  - Then we show that the difference $\frac{1}{s_n}\sum_{k=1}^n X_k I_{|X_k|>a_n} \xrightarrow{P} 0$.
  - Then S_n/s_n is asymptotically normal by Slutsky's theorem.

- Independence might not hold

  A basic remedy for sums of dependent random variables is to rewrite it as sum of independent random variables, with a negligible correction.

Normalizations that do not use the variance

Example 9. Let $X_1, X_2, \dots$ be independent random variables with the distribution ($k \ge 2$)

$$\begin{aligned}\Pr(X_k &= \pm 1) = 1/4,\\ \Pr(X_k &= k^2) = 1/k^2,\\ \Pr(X_k &= 0) = 1/2 - 1/k^2.\end{aligned}$$

Let $S_n = \sum_{k=2}^{n+1} X_k$. Then $E(X_k) = 1$ and $E(X_k^2) = \frac{1}{2} + k^2$ so $s_n^2 = \frac{1}{6}n\left(2n^2 - 3n + 4\right) \sim n^3/3$. One can check that $(S_n - n)/s_n \xrightarrow{P} 0$.

Because with a "proper normalization" and without any centering, we have $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} Z/\sqrt{2}$. To see this, note that $Y_k = X_k I_{|X|_k} \le 1$ are i.i.d. with mean 0, variance $\frac{1}{2}$ so their partial sums satisfy CLT.

Since $P(Y_k \ne X_k) = 1/k^2$ is a convergent series, by the first Borel Cantelli Lemma $|\frac{1}{\sqrt{n}}\sum_{k=1}^n (Y_k - X_k)| \le \frac{|\Sigma|}{\sqrt{n}} \to 0$ with probability one.

Example 10 (A good project for the final?). Suppose X_k are independent with the distribution

$$X_k = \begin{cases} 1 & \text{with probability } 1/2 - p_k \\ -1 & \text{with probability } 1/2 - p_k \\ k^\theta & \text{with probability } p_k \\ -k^\theta & \text{with probability } p_k \end{cases}$$

and $S_n = \sum_{k=1}^n X_k$. It is "clear" that if $\sum p_k < \infty$ then $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} N(0,\sigma^2)$ for any θ. It is "clear" that if $\theta = 0$ then $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} N(0,1)$ for any choice of $p_k < 1/2$.

So it is natural to ask what assumptions on θ and p_k will imply asymptotic normality. In paricular,

- What are the "optimal" restrictions on p_k if $\theta < 0$? (Say, if $\theta = -1$, to ease the calculations)
- Can one "do better" than $\sum p_k < \infty$ if $\theta > 0$? (Say, if $\theta = 1$, to ease the calculations)

CLT without independence

Example 11. Suppose ξ_k are i.i.d. with mean zero and variance $\sigma^2 > 0$. Show that the sums of moving averages $X_k = \frac{1}{m+1}\sum_{j=k}^{k+m} \xi_j$ satisfy the Central Limit Theorem.

Proof. Write $S_n = \sum_{k=1}^n X_k$. We will show that $\frac{1}{\sqrt{n}} S_n \xrightarrow{\mathcal{D}} N(0,\sigma^2)$.

$$S_n = \sum_{k=1}^n \frac{1}{m+1}\sum_{j=k}^{k+m} \xi_j = \sum_{j=1}^{n+m} \xi_j \sum_{k=1\vee(j-m)}^{n\wedge j} \frac{1}{m+1} = \sum_{j=1}^n \xi_j + R_n.$$

$$R_n = -\sum_{j=1}^m \frac{m+1-j}{m+1}\xi_j + \sum_{j=n+1}^{n+m} \frac{n+m+1-j}{m+1}\xi_j$$

By CLT for i.i.d random variables, $\frac{1}{\sigma\sqrt{n}}\sum_{j=1}^n \xi_j \xrightarrow{\mathcal{D}} N(0,1)$. So we only need to look more carefully at

Since $E(R_n^2) \le 2m^2\sigma^2$, we see that $R_n/\sqrt{n} \xrightarrow{P} 0$ so by Slutsky's theorem we get CLT. ◻

Example 12 (A good project for the final?). Suppose ξ_k are i.i.d. with mean zero and variance 1. Do "geometric moving averages"

$$X_k = \sum_{j=0}^{k} q^j \xi_{k-j}$$

satisfy the CLT when $|q| < 1$? That is, with $S_n = \sum_{k=1}^n X_k$ do we have $(S_n - a_n)/b_n \xrightarrow{\mathcal{D}} N(0,1)$ for appropriate normalizing constants a_n, b_n? And if so, how does b_n depend on the q?

Random normalizations

Example 13. Suppose $X_1, X_2, \dots,$ are i.i.d. with mean 0 and variance $\sigma^2 > 0$. Then

$$\frac{\sum_{k=1}^n X_k}{\sqrt{\sum_{k=1}^n X_k^2}}$$

converges in distribution to $N(0,1)$. To see this, write

$$\frac{\sum_{k=1}^n X_k}{\sqrt{\sum_{k=1}^n X_k^2}} = \frac{\sigma}{\sqrt{\frac{1}{n}\sum_{k=1}^n X_k^2}} \times \frac{\sum_{k=1}^n X_k}{\sigma\sqrt{n}}$$

and note that the first factor converges to 1 with probability one. To apply Slutsky's theorem, we now need to do some more work that is similar to some old exercises.

Writing $Z_n = \frac{\sigma}{\sqrt{\frac{1}{n}\sum_{k=1}^n X_k^2}}$, we check that $(Z_n - 1)S_n/\sqrt{\sigma^2 n} \xrightarrow{P} 0$. Choose arbitrary $\varepsilon > 0, K > 0$. Then $\limsup_{n\to\infty} P(|Z_n - 1| \cdot |S_n|/\sqrt{\sigma^2 n} > \varepsilon) \le \limsup_{n\to\infty} P(|S_n|/\sqrt{\sigma^2 n} > K) + \limsup_{n\to\infty} P(|Z_n - 1| > \varepsilon/K) \le \frac{1}{K^2}$. Since K is arbitrarily large, the limit it 0.

CLT without second moments

Exercise 1 (Exercise 11.5 from the notes). *Suppose X_k are independent and have density $\frac{1}{|x|^3}$ for $|x| > 1$. Show that $\frac{S_n}{\sqrt{n\log n}} \to N(0,1)$.*

Hint: *Verify that Lyapunov's condition* (7) *holds with $\delta = 1$ for truncated random variables.*

Solution

Let $Y_{nk} = X_k I_{|X_k| \le \sqrt{n}}$. Then $E(Y_{nk}) = 0$ by symmetry. Next we compute the variances

$$E(Y_{nk}^2) = 2\int_1^{\sqrt{n}} \frac{x^2}{x^3}\,dx == 2\int_1^{\sqrt{n}} \frac{dx}{x} = 2\log\sqrt{n} = \log n$$

Therefore $s_n^2 = \sum_{k=1}^n E(Y_{nk}^2) = n\log n$. To verify Lyapunov's condition (7) we compute $E(|Y_{nk}|^3) = 2\int_1^{\sqrt{n}} 1dx = 2\sqrt{n}$. This gives

$$\frac{1}{s_n^3}\sum_{k=1}^n E(|Y_{nk}|^3) = \frac{2n\sqrt{n}}{n\sqrt{n}\log n\sqrt{\log n}} = \frac{2}{(\log n)^{3/2}} \to 0$$

By Lyapunov's theorem (Theorem 6), we see that

$$\frac{1}{\sqrt{n\log n}}\sum_{k=1}^n Y_{nk} \xrightarrow{\mathcal{D}} N(0,1).$$

To finish the proof, we need to show that $\frac{1}{\sqrt{n\log n}}\sum_{k=1}^n Y_{nk} - \frac{1}{\sqrt{n\log n}}\sum_{k=1}^n X_k \xrightarrow{P} 0$. We show L_1-convergence. $E|Y_{kn} - X_k| = 2\int_{\sqrt{n}}^\infty x\frac{1}{x^3}\,dx = 2/\sqrt{n}$ so

$$E\left|\frac{1}{\sqrt{n\log n}}\sum_{k=1}^n Y_{nk} - \frac{1}{\sqrt{n\log n}}\sum_{k=1}^n X_k\right| \le \frac{1}{\sqrt{n\log n}}\sum_{k=1}^n E|X_k - Y_{nk}| \le \frac{2}{\sqrt{\log n}} \to 0$$

Exercise 2 (A good project for the final?). *Suppose X_k are i.i.d. with density $\frac{1}{|x|^3}$ for $|x| > 1$. Show that $\frac{S_n}{\sqrt{n\log n}} \to N(0,1)$ using one of the other truncations from the hint for Exercise 11.5 in the notes.*

Limit Theorems in $\mathbb{R}^k$ This is based on [Billingsley, Section 29].
April 24, 2020

4 The basic theorems

Notation

- If $\mathbf{X} : \Omega \to \mathbb{R}^k$ is measurable, then $\mathbf{X}$ is called a random vector. $\mathbf{X}$ is also called a k-variate random variable, as $\mathbf{X} = (X_1, \dots, X_k)$. We will also write $\mathbf{X}$ as column vectors.

- Recall that a probability distribution of $\mathbf{X}$ is a probability measure μ on Borel subsets of $\mathbb{R}^k$ defined by $\mu(U) = P(\{\omega : \mathbf{X}(\omega) \in U\})$.
- Recall that a (joint) cumulative distribution function of $\mathbf{X} = (X_1, \dots, X_n)$ is a function $F : \mathbb{R}^k \to [0,1]$ such that

$$F(x_1, \dots, x_k) = P(X_1 \le x_1, \dots, X_k \le x_k)$$

- From $\pi - \lambda$ theorem we know that F determines uniquely μ. In particular, if

$$F(x_1, \dots, x_k) = \int_{-\infty}^{x_1} \cdots \int_{-\infty}^{x_k} f(y_1, \dots, y_k) dy_1 \dots dy_k$$

  then $\mu(U) = \int_U f(y_1, \dots, y_k) dy_1 \dots dy_k$.

Let $\mathbf{X}_n : \Omega \to \mathbb{R}^k$ be a sequence of random vectors.

Definition 14. We say that $\mathbf{X}_n$ *converges in distribution* to $\mathbf{X}$ if for every bounded continuous function $f : \mathbb{R}^k \to \mathbb{R}$ the sequence of numbers $\mathbb{E}(f(\mathbf{X}_n))$ converges to $\mathbb{E} f(\mathbf{X})$.

We will write $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$; if μ_n is the law of $\mathbf{X}_n$ we will also write $\mu_n \xrightarrow{\mathcal{D}} \mu$; the same notation in the language of cumulative distribution functions is $F_n \xrightarrow{\mathcal{D}} F$; the latter can be defined as $F_n(\mathbf{x}) \xrightarrow{\mathcal{D}} F(\mathbf{x})$ for all points of continuity of F, but it is simpler to use Definition 14.

Proposition 15. *If $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ and $g : \mathbb{R}^k \to \mathbb{R}^m$ is a continuous function then $g(\mathbf{X}_n) \xrightarrow{\mathcal{D}} g(\mathbf{X})$*

For example, if $(X_n, Y_n) \xrightarrow{\mathcal{D}} (Z_1, Z_2)$ then $X_n^2 + Y_n^2 \xrightarrow{\mathcal{D}} Z_1^2 + Z_2^2$.

Proof. Denoting by $\mathbf{Y}_n = g(\mathbf{X}_n)$, we see that for any bounded continuous function $f : \mathbb{R}^m \to \mathbb{R}$, $f(\mathbf{Y}_n)$ is a bounded continuous function $f \circ g$ of $\mathbf{X}_n$. □

Definition 16. The sequence of measures μ_n on $\mathbb{R}^k$ is tight if for every $\varepsilon > 0$ there exists a compact set $K \subset \mathbb{R}^k$ such that $\mu_n(K) \ge 1 - \varepsilon$ for all n.

Theorem 17. *If μ_n is a tight sequence of probability measures then there exists μ and a subsequence n_k such that $\mu_{n_k} \xrightarrow{\mathcal{D}} \mu$*

Proof. The detailed proof is omitted. Omitted in 2020 □

Corollary 18. *If $\{\mu_n\}$ is a tight sequence of probability measures on Borel subsets of $\mathbb{R}^k$ and if each convergent subsequence has the same limit μ, then $\mu_n \xrightarrow{\mathcal{D}} \mu$*

The end
Lets stop here

- Things to do on Monday:
  - Multivariate characteristic functions and multivariate normal distribution.

Thank you

5 Multivariate characteristic function

Multivariate characteristic function and multivariate normal distribution

Multivariate characteristic function

Recall the dot-product $\mathbf{x} \cdot \mathbf{y} := \mathbf{x}'\mathbf{y} = \sum_{j=1}^{k} x_j y_j$.

- The multivariate characteristic function $\varphi : \mathbb{R}^k \to \mathbb{C}$ is

$$\varphi(\mathbf{t}) = \mathbb{E} \exp(i\mathbf{t} \cdot \mathbf{X}) \tag{8}$$

- This is also written as $\varphi(t_1, \ldots, t_k) = E \exp(\sum_{j=1}^{k} it_j X_j)$.
- The inversion formula shows how to determine $\mu(U)$ for a rectangle $U = (a_1, b_1] \times (a_2, b_2] \times \cdots \times (a_k, b_k]$ such that $\mu(\partial U) = 0$:

$$\mu(U) = \lim_{T \to \infty} \frac{1}{(2\pi)^k} \int_{-T}^{T} \cdots \int_{-T}^{T} \prod_{j=1}^{k} \frac{e^{-ia_k j t_j} - e^{-ib_j t_j}}{it_j} \varphi(t_1, \ldots, t_k) dt_1 \ldots dt_k \tag{9}$$

- Thus the characteristic function determines the probability measure μ uniquely.

Corollary 19 (Cramer-Wold device I). *The law of* $\mathbf{X}$ *is uniquely determined by the univariate laws* $\mathbf{t} \cdot \mathbf{X} = \sum_{j=1}^{k} t_j X_j$.

Corollary 20. X, Y *are independent iff* $\varphi_{X,Y}(s,t) = \varphi_X(s)\varphi_Y(t)$

Example 21. If X, Y are independent normal with the same variance then $X+Y$ and $X-Y$ are independent normal. Indeed, WLOG we assume that means are zero and variances are one. $\varphi_{X+Y,X-Y}(s,t) = \mathbb{E}e^{is(X+Y)+it(X-Y)} = \mathbb{E}e^{i(t+s)X+i(s-t)Y} = \varphi_X(s+t)\varphi_Y(s-t) = \exp((t+s)^2/2 + (s-t)^2/2) = \exp((t^2+s^2+2ts)/2 + (s^2+t^2-2st)/2) = e^{s^2}e^{t^2}$. This matches $\varphi_{X+Y}(s)\varphi_{X-Y}(t)$ as $\varphi_{X\pm Y}(s) = e^{s^2/2}e^{s^2/2} = e^{s^2}$.

Theorem 22 (Bernstein (1941)). *If* X, Y *are independent and* $X+Y, X-Y$ *are independent, then* X, Y *are normal*

Kac M. "On a characterization of the normal distribution," American Journal of Mathematics. **1939**. 61. pp. 726—728.

Theorem 23 (Cramer-Wold device II). $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$ *iff* $\varphi_n(\mathbf{t}) \to \varphi(\mathbf{t})$ *for all* $\mathbf{t} \in \mathbb{R}^k$.

Note that this means that $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$ iff for all $t_1, \ldots, t_k$ univariate random variables converge, $\sum t_j X_j(n) \xrightarrow{\mathcal{D}} \sum t_j Y_j$

Corollary 24. *If* $Z_1, \ldots Z_m$ *are independent normal,* $\mathbf{A}$ *is an* $k \times m$ *matrix and* $\mathbf{X} = \mathbf{AZ}$ *then* $\sum_{j=1}^{k} t_j X_j$ *is (univariate) normal.*

Proof. Lets simplify the calculations by assuming Z_j are standard normal. The characteristic function of $S = \sum_j t_j X_j$ is

$$\varphi(s) = E\exp(is(\mathbf{t} \cdot \mathbf{X})) = E\exp(is(\mathbf{t} \cdot \mathbf{AZ})) = E\exp(is(\mathbf{A}'\mathbf{t}) \cdot \mathbf{Z})$$

$$= \prod_{i=1}^{k} e^{-s^2[\mathbf{A}^T\mathbf{t}]_i^2/2} = e^{-s^2\|\mathbf{A}'\mathbf{t}\|^2/2}$$

So S is $N(0, \sigma^2)$ with variance $\sigma^2 = \|\mathbf{A}'\mathbf{t}\|^2$ □

The generalization of this property is the "cleanest" definition of the mutlivariate normal distribution.

6 Multivariate normal distribution

Multivariate normal distribution $N(\mathbf{m}, \Sigma)$

Definition 25. $\mathbf{X}$ is *multivariate normal* if there is a vector $\mathbf{m}$ and a positive-definite matrix Σ such that its characteristic function is

$$\varphi(\mathbf{t}) = \exp\left(i\mathbf{t}'\mathbf{m} - \tfrac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right) \tag{*}$$

(How do we know that (*) n is a characteristic function?) By differentiation $\frac{\partial}{\partial t_j}$ and $\frac{\partial^2}{\partial t_i \partial t_j}$, the parameters $N(\mathbf{m}, \Sigma)$ get natural interpretation: $\mathbb{E}\mathbf{X} = \mathbf{m}$ and $\Sigma_{i,j} = \mathrm{cov}(X_i, X_j)$ so $\Sigma = \mathbb{E}(\mathbf{X}\mathbf{X}') - \mathbf{m}\mathbf{m}'$.

Definition 26. $\mathbf{X}$ is *multivariate normal* if there is a vector $\mathbf{m}$ an $m \times k$ matrix $\mathbf{A}$ and a sequence $Z_1, \dots, Z_m$ of independent standard normal random variables such that $\mathbf{X} = \mathbf{m} + \mathbf{A}\mathbf{Z}$

Note that previous slide says $\varphi_{\mathbf{t}'(\mathbf{X}-\mathbf{m})}(s) = e^{-s^2\|\mathbf{A}'\mathbf{t}\|^2/2}$ shows that $\mathbf{X}$ has characteristic function (*) and $\mathbf{t} \cdot \mathbf{X}$ has variance

$$\sigma^2 = \|\mathbf{A}'\mathbf{t}\|^2 = (\mathbf{A}'\mathbf{t}) \cdot (\mathbf{A}'\mathbf{t}) = \mathbf{t}'\mathbf{A}\mathbf{A}'\mathbf{t} = \mathbf{t}'\Sigma\mathbf{t}$$

If $\mathbf{m} = 0$ then $\mathbb{E}\mathbf{X}\mathbf{X}' = \mathbb{E}\mathbf{A}\mathbf{Z}\mathbf{Z}'\mathbf{A}' = \mathbf{A}\mathbb{E}(\mathbf{Z}\mathbf{Z}')\mathbf{A}' = \mathbf{A}\mathbf{A}' = \Sigma$

Definition 27. $\mathbf{X}$ is *multivariate normal* if for every $\mathbf{t} \in \mathbb{R}^k$ the univariate random variable $X = \mathbf{X} \cdot \mathbf{t}$ is normal $N(\mu, \sigma^2)$ for some $\mu = \mu(\mathbf{t}) \in \mathbb{R}$ and $\sigma^2 = \sigma^2(\mathbf{t}) \geq 0$.

Multivariate normal distribution $N(\mathbf{m}, \Sigma)$

Remark 28. *If $\mathbf{X}$ is normal $N(\mathbf{m}, \Sigma)$, then $\mathbf{X} - \mathbf{m}$ is centered normal $N(0, \Sigma)$. In the sequel, to simplify notation we only discuss centered case.*

Here is the fourth definition:

Definition 29 (half-definition)**.** $\mathbf{X}$ is $N(0, \Sigma)$ if it has density

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{k/2}\sqrt{\det \Sigma}} \exp\left(-\frac{\mathbf{x} \cdot (\Sigma^{-1}\mathbf{x})}{2}\right)$$

We are not going to use this definition!

Remark 30. *Denoting by $\mathbf{a}_k$ the columns of $\mathbf{A}$, we have $\mathbf{X} = \sum_{j=1}^k Z_j \mathbf{a}_j$. This is the universal feature of Gaussian vectors, even in infinite-dimensional vector spaces – they all can be written as linear combinations of deterministic vectors with independent real-valued "noises" as coefficients. For example, the random "vector" $(W_t)_{0 \leq t \leq 1}$ with values in the vector space $C[0,1]$ of continuous functions on $[0,1]$ can be written as $W_t = \sum_{k=1}^\infty Z_j g_j(t)$ with deterministic functions $g_j(t) = \frac{1}{2j+1}\sin((2j+1)\pi t)$.*

Example: bivariate $N(0, \Sigma)$

- Write $\mathbf{X} = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$. WLOG assume $\mathbb{E}(X_1) = \mathbb{E}(X_2) = 0$ and $\mathbb{E}(X_1^2) = \mathbb{E}(X_2^2) = 1$. Then there is just one free parameter: correlation coefficient $\rho = \mathbb{E}(X_1X_2)$.
- $\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ is non-negative definite for any $|\rho| \leq 1$ and $\varphi(s,t) = e^{-s^2/2 - t^2/2 - \rho st}$ is a characteristic function of a random variable $\mathbf{X} = (X_1, X_2)$ with univariate N(0,1) laws, with correlation $\mathbb{E}(X_1X_2) = -\frac{\partial^2}{\partial s \partial t}\varphi(s,t)|_{s=t=0} = \rho$.
- If Z_1, Z_2 are independent $N(0,1)$ then

$$X_1 = Z_1, \; X_2 = \rho Z_1 + \sqrt{1-\rho^2} Z_2 \tag{10}$$

  will have exactly the same second moments, and the same characteristic function.
- Since $\det \Sigma = 1 - \rho^2$, when $\rho^2 \neq 1$ the matrix is invertible and the resulting bivariate normal density is

$$f(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{x^2 + y^2 - 2\rho xy}{2(1-\rho^2)}\right)$$

- From (10) we also see that $X_2 - \rho X_1$ is independent of X_1 and has variance $1 - \rho^2$. In particular if $\rho = 0$ then X_1, X_2 are independent.

Remark 31. *The covariance matrix $\Sigma = \mathbf{A}\mathbf{A}'$ is unique but the representation $\mathbf{X} = \mathbf{A}\mathbf{Z}$ is not unique. For example independent pair*

$$\mathbf{X} = \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \times \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

can also be represented as

$$\mathbf{X} = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ 1/\sqrt{2} & -1/\sqrt{2} \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix}$$

because $Z_1 - Z_2$ and $Z_1 + Z_2$ are independent normal random variables of variance 2 and $\tilde{\mathbf{X}} = \begin{bmatrix} (Z_1 + Z_2)/\sqrt{2} \\ (Z_1 - Z_2)/\sqrt{2} \end{bmatrix}$ has the same law as $\mathbf{X}$. This implies non-uniqueness for all other representations.

Normal distributions on octonions

(I do not know the answer for octonions)

Example 32 (Good project for the final?). Suppose Z_1, Z_2, Z_3, Z_4 be independent normal random variables. Let $\mathbf{Z}_{\mathbb{C}} = Z_1 + iZ_2$ be a complect random variable and $\mathbf{Z}_{\mathbb{Q}} = Z_1 + iZ_2 + jZ_3 + kZ_4$ be a quaternionic random variable.

- Show that
$$\mathbb{E}Z_1^n = \begin{cases} \frac{n!}{2^{n/2}(n/2)!} & \text{if } n \text{ is even} \\ 0 \end{cases}$$
- What is the formula for $\mathbb{E}(\mathbf{Z}_{\mathbb{C}}^n)$ and for $\mathbb{E}(\mathbf{Z}_{\mathbb{C}}^m \bar{\mathbf{Z}}_{\mathbb{C}}^n)$ for $m, n = 0, 1, 2, \ldots$?
- What is the formula for $\mathbb{E}(\mathbf{Z}_{\mathbb{Q}}^n)$ and for $\mathbb{E}(\mathbf{Z}_{\mathbb{Q}}^m \bar{\mathbf{Z}}_{\mathbb{Q}}^n)$ for $m, n = 0, 1, 2, \ldots$

These are questions about gaussian random matrices $\mathbf{Z}_{\mathbb{C}} = \begin{bmatrix} Z_1 & Z_2 \\ -Z_2 & Z_1 \end{bmatrix}$ and $\mathbf{z}_{\mathbb{Q}} = \begin{bmatrix} \mathbf{z}_{\mathbb{C}} & \tilde{\mathbf{z}}_{\mathbb{C}} \\ -\tilde{\mathbf{z}}_{\mathbb{C}}^T & \mathbf{z}_{\mathbb{C}}^T \end{bmatrix} = \begin{bmatrix} Z_1 & Z_2 & Z_3 & Z_4 \\ -Z_2 & Z_1 & -Z_4 & Z_3 \\ -Z_3 & Z_4 & Z_1 & -Z_2 \\ -Z_4 & -Z_3 & Z_2 & Z_1 \end{bmatrix}$

The end

Lets stop here

- Things to do on Wednesday:
  - Multivariate central limit theorem.
  - Examples
  - Final Exam projects

Thank you

Mutlivariate CLT and applications

Recall from previous lectures

- The multivariate characteristic function $\varphi(\mathbf{t}) = \mathbb{E}\exp(i\mathbf{t} \cdot \mathbf{X})$
- This is also written as $\varphi(\mathbf{t}) = \mathbb{E}\exp(i\mathbf{t}'\mathbf{X})$.
- This is also written as $\varphi(t_1, \ldots, t_k) = E\exp(\sum_{j=1}^k it_j X_j)$.

Theorem 33 (Cramer-Wold device II). $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$ *iff* $\varphi_n(\mathbf{t}) \to \varphi(\mathbf{t})$ *for all* $\mathbf{t} \in \mathbb{R}^k$.

Note that this means that $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{Y}$ iff for all $t_1, \ldots, t_k$ univariate random variables converge, $\sum t_j X_j(n) \xrightarrow{\mathcal{D}} \sum t_j Y_j$

Definition 34. $\mathbf{X}$ is *multivariate normal* if there is a vector $\mathbf{m}$ and a positive-definite matrix Σ such that its characteristic function is $\varphi(\mathbf{t}) = \exp\left(i\mathbf{t}'\mathbf{m} - \frac{1}{2}\mathbf{t}'\Sigma\mathbf{t}\right)$.

Equivalently, $\mathbf{X} = \mathbf{m} + \mathbf{A}\mathbf{Z}$, where $\mathbf{A}\mathbf{A}' = \Sigma$. Without loss of generality we can assume $\mathbf{A}$ is a square matrix.

Equivalently, $\mathbf{X} = \mathbf{m} + \sum_{j=1}^k \vec{v}_j Z_j$, where Z_j are i.i.d. $N(0,1)$ and $\Sigma = \sum_{j=1}^k \vec{v}_j \vec{v}_j'$.

7 The CLT

The CLT

Theorem 35. *Let $\mathbf{X}_n = (X_{n1}, \dots, X_{nk})$ be independent random vectors with the same distribution and finite second moments. Denote $\mathbf{m} = E\mathbf{X}_k$ and $\mathbf{S}_n = \mathbf{X}_1 + \dots + \mathbf{X}_n$. Then*

$$(\mathbf{S}_n - n\mathbf{m})/\sqrt{n} \xrightarrow{\mathcal{D}} \mathbf{Y}$$

where $\mathbf{Y}$ is a centered normal distribution with the covariance matrix $\Sigma = E(\mathbf{X}_n\mathbf{X}_n') - \mathbf{m}\mathbf{m}'$.

The notation is $N(0, \Sigma)$. Note that this is inconsistent with the univariate notation $N(\mu, \sigma)$ which for consistency with the multivariate case should be replaced by $N(\mu, \sigma^2)$.

Proof. Without loss of generality we can assume $\mathbf{m} = 0$. Let $\mathbf{t} \in \mathbb{R}^k$. Then $X_n := \mathbf{t}'\mathbf{X}_n$ are univariate i.i.d. variables with mean zero and variance $\sigma^2 = \mathbb{E}(\mathbf{t}'\mathbf{X}_n)^2 = \mathbb{E}(\mathbf{t}'\mathbf{X}_n\mathbf{X}_n'\mathbf{t}) = \mathbf{t}'\mathbb{E}(\mathbf{X}_n\mathbf{X}_n')\mathbf{t} = \mathbf{t}'\Sigma\mathbf{t}$. By CLT for i.i.d. case, we have $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} \sigma Z$.

If $\mathbf{Y} = (Y_1, \dots, Y_k)$ has multivariate normal distribution with covariance Σ, then $\mathbf{t}'\mathbf{Y}$ is univariate normal with the same variance σ^2. So we showed that $\mathbf{t}'\mathbf{S}_n/\sqrt{n} \xrightarrow{\mathcal{D}} \mathbf{t}'\mathbf{Y}$ for all $\mathbf{t} \in \mathbb{R}^k$. This ends the proof by Theorem 33 (Cramer-Wold device).

$\square$

Example 36. Suppose ξ_k, η_k are i.i.d with mean zero variance one. Then $\frac{1}{\sqrt{n}}(\sum_{k=1}^n \eta_k, \sum_{k=1}^n(\eta_k + \xi_k)) \xrightarrow{\mathcal{D}} N(0, \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix})$.

Indeed, random vector $\mathbf{X}_k = \begin{bmatrix} \xi_k \\ \xi_k + \eta_k \end{bmatrix}$ has covariance matrix $\Sigma = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$

Since $\sum_{k=1}^n \mathbf{X}_k = \begin{bmatrix} S_n^\eta \\ S_n^\eta \end{bmatrix} + \begin{bmatrix} 0 \\ S_n^\xi \end{bmatrix}$, this is not anything impressive, as $\begin{bmatrix} Z_1 \\ Z_1 + Z_2 \end{bmatrix}$ has the required covariance matrix.

7.1 Application: Chi-Squared test for multinomial distribution

Application

Chi-Squared test for multinomial distribution

- A multinomial experiment has k outcomes with probabilities $p_1, \dots, p_k > 0$.
- A multinomial random variable $(N_1, \dots, N_k)$ lists observed counts per category in n repeats of the multinomial experiment. The expected counts are then $E_j = np_j$.
- The following result is behind the use of the chi-squared statistics in tests of consistency.

Theorem 37. $\sum_{j=1}^k \frac{(N_j - E_j)^2}{E_j} \xrightarrow{\mathcal{D}} \chi^2_{k-1} = Z_1^2 + \dots + Z_{k-1}^2$

Lets write this in our language: take i.i.d. vectors $P(\mathbf{X} = \vec{e}_j) = p_j$ and let $\mathbf{S}(n) = \sum_{j=1}^n \mathbf{X}_j$. Then

Theorem 38. $\sum_{j=1}^k \frac{(S_j(n) - np_j)^2}{np_j} \xrightarrow{\mathcal{D}} Z_1^2 + \dots + Z_{k-1}^2$

$\sum_{j=1}^k \frac{(S_j(n) - np_j)^2}{np_j} \xrightarrow{\mathcal{D}} Z_1^2 + \dots + Z_{k-1}^2$

Lets prove this for $k = 3$. Consider independent random vectors $\mathbf{X}_k$ that take three values $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$ and $\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$ with probabilities p_1, p_2, p_3. Then $\mathbf{S}_n$ is the sum of n independent identically distributed vectors $\mathbf{X}_1, \dots, \mathbf{X}_n$. Components of $\mathbf{S}_n$ are counts

Clearly, $E\mathbf{X}_k = \begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}$. To compute the covariance matrix, write $\mathbf{X}$ for $\mathbf{X}_k$. For non-centered vectors, the covariance is $E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X}')$. We have

$$E(\mathbf{X}\mathbf{X}') = p_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + p_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \times \begin{bmatrix} 0 & 1 & 0 \end{bmatrix} + p_3 \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \times \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} p_1 & 0 & 0 \\ 0 & p_2 & 0 \\ 0 & 0 & p_3 \end{bmatrix}$$

So

$$\Sigma = E(\mathbf{X}\mathbf{X}') - E(\mathbf{X})E(\mathbf{X}') = \begin{bmatrix} p_1(1-p_1) & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & p_2(1-p_2) & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & p_3(1-p_3) \end{bmatrix}$$

Then $\mathbf{S}_n$ is the sum of n independent vectors, and the central limit theorem implies that $\frac{1}{\sqrt{n}}\left(\mathbf{S}_n - n\begin{bmatrix} p_1 \\ p_2 \\ p_3 \end{bmatrix}\right) \xrightarrow{\mathcal{D}} \mathbf{W}$. By Continuity Theorem 15 we have

$$\sum_{j=1}^{3} \frac{(S_j(n)-np_j)^2}{np_j} \xrightarrow{\mathcal{D}} \sum_{j=1}^{3} \frac{W_j^2}{p_j}$$

where $\mathbf{W} = (W_1, W_2, W_3)$ is multivariate normal with covariance matrix Σ.

$\mathbf{W}$ is $N(0,\Sigma)$

Note that since $\sum_{j=1}^k S_j(n) = n$, the gaussian distribution is degenerate: $W_1 + W_2 + W_3 = 0$. (No density!)

It remains to show that $\sum_{j=1}^3 \frac{W_j^2}{p_j}$ has the same law as $Z_1^2 + Z_2^2$ i.e. that it is exponential. To do so, we first note that the covariance of $(Y_1, Y_2, Y_3,) := (W_1/\sqrt{p_1}, W_2/\sqrt{p_2}, W_3/\sqrt{p_3})$ is

$$\Sigma_{\mathbf{Y}} = \begin{bmatrix} 1-p_1 & -\sqrt{p_1p_2} & -\sqrt{p_1p_3} \\ -\sqrt{p_1p_2} & 1-p_2 & -\sqrt{p_2p_3} \\ -\sqrt{p_1p_3} & -\sqrt{p_2p_3} & 1-p_3 \end{bmatrix} = I - \begin{bmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{bmatrix} \times \begin{bmatrix} \sqrt{p_1} & \sqrt{p_2} & \sqrt{p_3} \end{bmatrix}$$

Since $\mathbf{v}_1 = \begin{bmatrix} \sqrt{p_1} \\ \sqrt{p_2} \\ \sqrt{p_3} \end{bmatrix}$ is a unit vector, we can complete it with two additional vectors $\mathbf{v}_2 = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix}$ and $\mathbf{v}_3 = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$ to form an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$. This can be done in many ways, for example by the Gram-Schmidt orthogonalization to $\mathbf{v}_1$, $\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$. The specific form of $\mathbf{v}_2, \mathbf{v}_3$ does not enter the proof - we only need to know that $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$ are *orthonormal*.

$\Sigma_{\mathbf{Y}} = I - \mathbf{v}_1\mathbf{v}_1'$

To complete the proof we write $I = \mathbf{v}_1\mathbf{v}_1' + \mathbf{v}_2\mathbf{v}_2' + \mathbf{v}_3\mathbf{v}_3'$ as these are orthogonal eigenvectors of I with $\lambda = 1$. (Or, because $\mathbf{x} = \mathbf{v}_1\mathbf{v}_1'\mathbf{x} + \mathbf{v}_2\mathbf{v}_2'\mathbf{x} + \mathbf{v}_3\mathbf{v}_3'\mathbf{x}$ as $\mathbf{v}_j'\mathbf{x} = \mathbf{x}\cdot\mathbf{v}_j$ are the coefficients of expansion of $\mathbf{x}$ in orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of $\mathbb{R}^3$.)

Therefore,

$$\Sigma_{\mathbf{Y}} = \mathbf{v}_2\mathbf{v}_2' + \mathbf{v}_3\mathbf{v}_3'$$

We now notice that $\Sigma_{\mathbf{Y}}$ is the covariance of another multivariate normal random variable $\mathbf{Z} = \mathbf{v}_2 Z_2 + \mathbf{v}_3 Z_3$ where Z_2, Z_3 are independent real-valued $N(0,1)$. Indeed,

$$E\mathbf{Z}\mathbf{Z}' = \sum_{i,j=2}^{3} \mathbf{v}_i\mathbf{v}_j' E(Z_iZ_j) = \sum_{i=2}^{3} \mathbf{v}_i\mathbf{v}_i' = \mathbf{v}_2\mathbf{v}_2' + \mathbf{v}_3\mathbf{v}_3'$$

Therefore, vector $\mathbf{Y}$ has the same distribution as $\mathbf{Z}$, and the square of its length $Y_1^2 + Y_2^2 + Y_3^2$ has the same distribution as

$$\|\mathbf{Z}\|^2 = \|\mathbf{v}_2 Z_2 + \mathbf{v}_3 Z_3\|^2 = \|\mathbf{v}_2 Z_2\|^2 + \|\mathbf{v}_3 Z_3\|^2 = Z_2^2 + Z_3^2$$

(recall that $\mathbf{v}_2$ and $\mathbf{v}_3$ are orthogonal unit vectors).

Remark 39 (Good project for the final?)**.** *It is clear that this proof generalizes to all k.*

The distribution of $Z_1^2 + \cdots + Z_{k-1}^2$ is Gamma with parameters $\alpha = (k-1)/2$ and $\beta = 2$, known in statistics as chi-squared distribution with $k-1$ degrees of freedom. To see that $Z_2^2 + Z_3^2$ is indeed chi-squared with two-degrees of freedom (i.e., exponential), we can determine the cumulative distribution function by computing $1 - F(u)$:

$$P(Z_2^2 + Z_3^2 > u) = \frac{1}{2\pi}\int_{x^2+y^2>u} e^{-(x^2+y^2)/2}dxdy$$

$$= \frac{1}{2\pi}\int_0^{2\pi}\int_{r>\sqrt{u}} e^{-r^2/2} r dr d\theta = e^{-u/2}$$

To compute the density of Z_1^2, differentiate $F_{Z_1^2}(x) = \frac{1}{\sqrt{2pi}}\int_{-\sqrt{x}}^{\sqrt{x}} e^{-z^2/2}dz$. These are cases $m=2$ and $m=1$ of the formula from Wikipedia:

$$f(x;\, m) = \begin{cases} \dfrac{x^{\frac{m}{2}-1}e^{-\frac{x}{2}}}{2^{\frac{m}{2}}\Gamma\left(\frac{m}{2}\right)}, & x > 0; \\ 0, & \text{otherwise.} \end{cases}$$

Example 40 (Good project for Final). Suppose ξ_j, η_j, γ_j are i.i.d. mean zero variance 1. Construct the following vectors:

$$\mathbf{X}_j = \begin{bmatrix} \xi_j - \eta_j \\ \eta_j - \gamma_j \\ \gamma_j - \xi_j \end{bmatrix}$$

Let $\mathbf{S}_n = \mathbf{X}_1 + \cdots + \mathbf{X}_n$. Show that $\frac{1}{n}\|\mathbf{S}_n\|^2 \xrightarrow{\mathcal{D}} Y$, and determine the density of Y.

Exercise 3 (Mutlivariate Slutsky's Thm). *Suppose that $\mathbb{R}^{2k}$-valued random variables $(\mathbf{X}_n, \mathbf{Y}_n)$ are such that $\mathbf{X}_n \xrightarrow{\mathcal{D}} \mathbf{X}$ and $\mathbf{Y}_n \xrightarrow{P} 0$ (that is, $\lim_{n\to\infty} P(\|\mathbf{Y}_n\| > \varepsilon) = 0$ for all $\varepsilon > 0$).*

Prove that $\mathbf{X}_n + \mathbf{Y}_n \xrightarrow{\mathcal{D}} \mathbf{X}$

The end

Lets stop here

- Things to do on Friday:
  - Questions?
  - Curiosities:
    - Iserlis theorem (Wick formula).
    - Wigner matrices
    - Wishart matrices
  - Final Exam projects

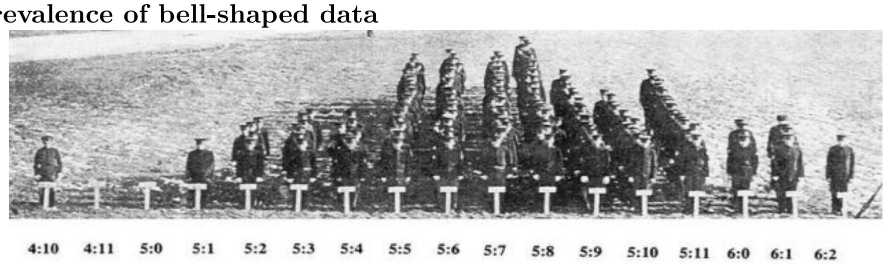
Thank you

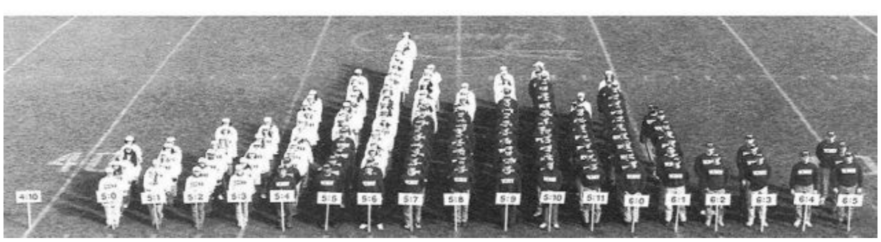
Additional topics

Today's plan

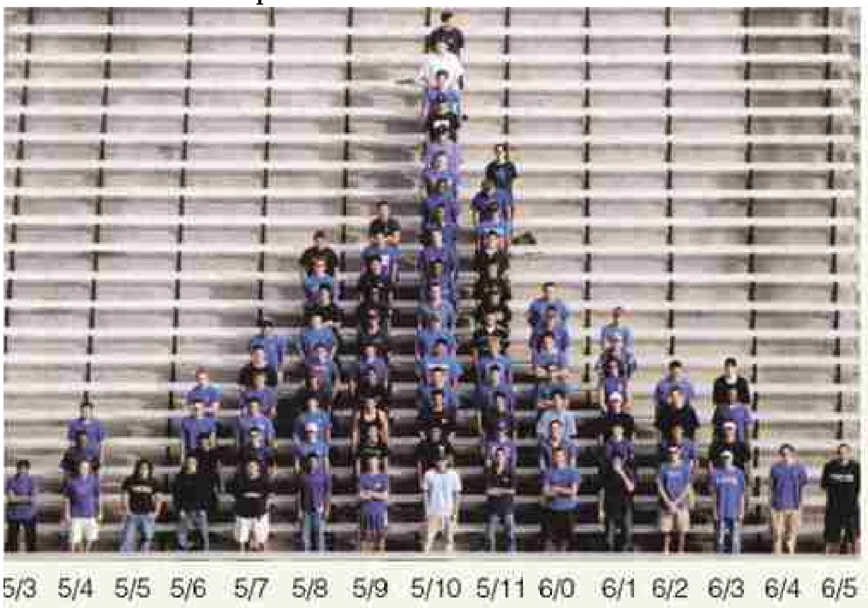
- Q&A
- Joint moments of multivariate normal distribution
- Random matrices

Prevalence of bell-shaped data

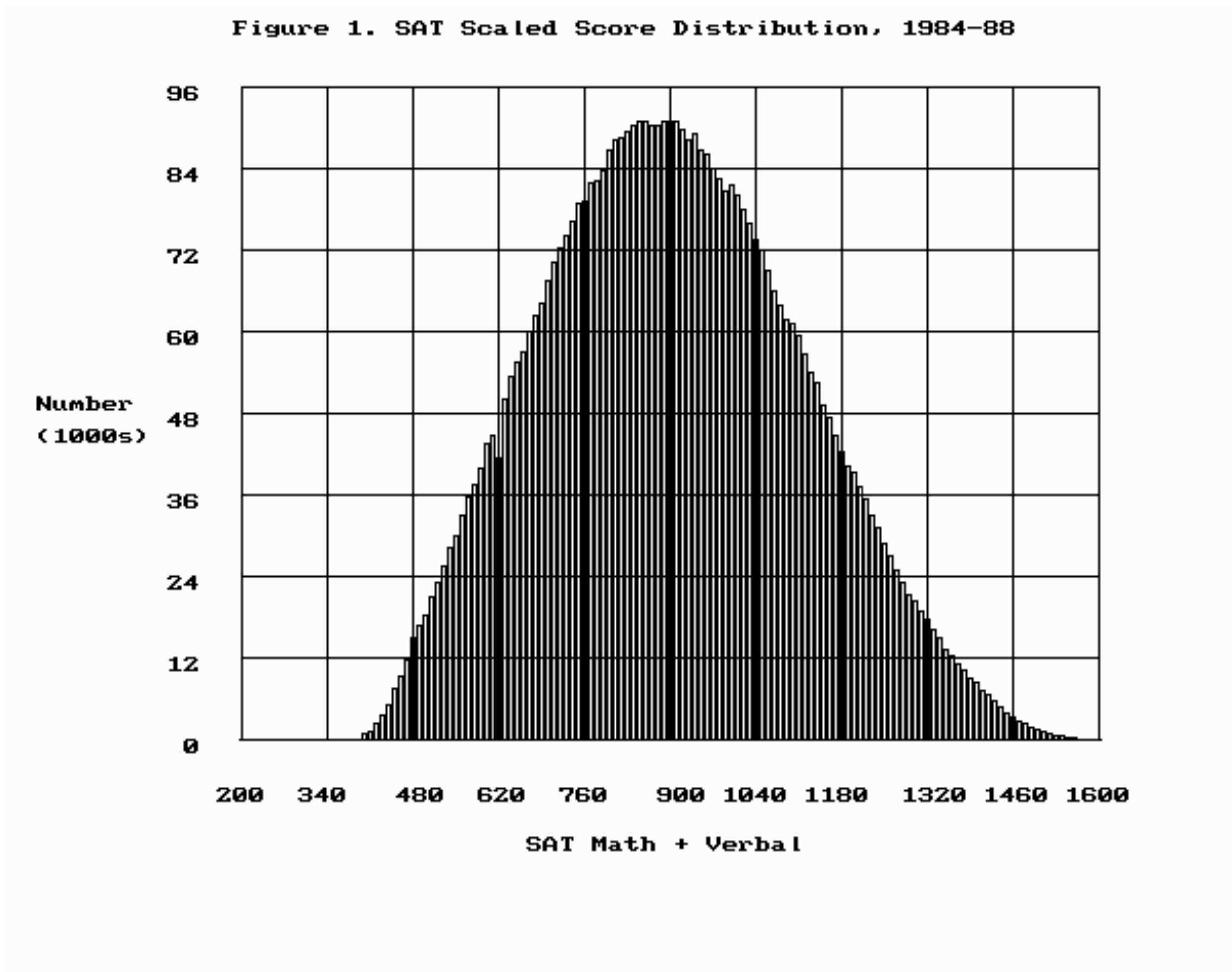




Prevalence of bell-shaped data



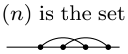
Prevalence of bell-shaped data



Theorem 41 (Isserlis (1918), Wick (1950))**.** *If* $\mathbf{X}$ *is* $N(0, \Sigma)$ *then*

$$\mathbb{E}(X_1 X_2 \dots X_k) = \sum_{\pi \in \mathcal{P}_2(k)} \prod_{\{i,j\} \in \pi} \mathbb{E}(X_i X_j)$$

Here $\mathcal{P}_2(n)$ is the set of pair partions of $\{1, \dots, k\}$. For example, there are three pair partitions for $\{1, 2, 3, 4\}$:

$\pi_1 = \{\{1,2\},\{3,4\}\}$, $\pi_2 = \{\{1,4\},\{2,3\}\}$, $\pi_3 = \{\{1,3\},\{2,4\}\}$.

So $E(X_1 X_2 X_3 X_4) = \Sigma_{1,2}\Sigma_{3,4} + \Sigma_{1,4}\Sigma_{2,3} + \Sigma_{1,3}\Sigma_{2,4}$. In particular,

- If Z is standard normal $\mathbb{E}(Z^4) = 3$ because we can apply the theorem to (Z, Z, Z, Z)

- If X, Y are jointly normal with variance 1 and correlation ρ then $\mathbb{E}(X^2Y^2) = 1 + 2\rho^2$ because we can apply the theorem to (X, X, Y, Y)
- If Z is standard normal then $E(Z^{2n}) = 1 \times 3 \times 5 \times \cdots \times (2n-1)$ because there are $2n-1$ choices to pair 1, then $2n-3$ choices to pair the next element on the list, and so on.

A 102 years ago ...
Isserlis, Biometrika (1918)

ON A FORMULA FOR THE PRODUCT-MOMENT COEFFICIENT OF ANY ORDER OF A NORMAL FREQUENCY DISTRIBUTION IN ANY NUMBER OF VARIABLES.

By L. ISSERLIS, D.Sc.

1. In *Biometrika*, Vol. XI, Part III, I have shown that for a normal frequency distribution in four variables, if

$$p_{xyzt} = \underset{x\ y\ z\ t}{SSSS} \{n_{xyzt}\, xyzt\}/N$$

denotes the product-moment coefficient of the distribution about the means of the four variables and q_{xyzt} is the *reduced* moment, i.e.

$$q_{xyzt} = p_{xyzt}/\sigma_x\sigma_y\sigma_z\sigma_t,$$

then
$$q_{xyzt} = r_{xy}r_{zt} + r_{yz}r_{xt} + r_{zx}r_{yt} \quad\text{.........................(1).}$$

In this result any two or more variables may be made identical leading to a variety of results for moment coefficients of distributions containing fewer than four variables but of total order four, for example identifying t with x we obtain

$$q_{x^2yz} = r_{yz} + 2r_{xy}r_{xz} \quad\text{...............................(2),}$$

and putting $y = z = t = x$ we find $q_{x^4} = 3$; of course $q_{xy} = r_{xy}$ and q_{x^4} is merely β_2.

I suggested that (1) was probably capable of generalisation, and I now propose to prove a general theorem which gives immediately the value of the mixed moment coefficient of any order in each variable for a normal frequency distribution in any number of variables.

2. Consider a normal distribution, total population N. Let $N_{12\ldots n}$ denote the frequency of the group in which the characters differ by $x_1, x_2, \ldots x_n$ from the mean values for the whole population and let

$$p_{1^{l_1}2^{l_2}\ldots n^{l_n}} = S\,(N_{12\ldots n}x_1^{l_1}x_2^{l_2}\ldots x_n^{l_n})/N \quad\text{...................(3),}$$

denote the moment coefficient of the most general kind about the mean values of the characters. The corresponding reduced moment will be

$$q_{1^{l_1}2^{l_2}\ldots n^{l_n}} = p_{1^{l_1}2^{l_2}\ldots n^{l_n}}/\sigma_1^{l_1}\sigma_2^{l_2}\ldots\sigma_n^{l_n} \quad\text{.......................(4).}$$

Then for *normal distributions*,

$$\text{if } n \text{ be } odd,\quad q_{12\ldots n} = 0 \quad\text{.................................(5),}$$

$$\text{and if } n \text{ be } even,\quad q_{12\ldots n} = S\,(r_{ab}r_{cd}\ldots r_{hk}) \quad\text{.................(6),}$$

where the summation on the right-hand side extends to every possible selection of $n/2$ pairs $ab, cd, \ldots hk$, that can be formed out of the n suffixes $1, 2, 3, \ldots n$; equation (1) is thus a particular case of (6).

Equation (6) is the theorem it is proposed to prove. The value of $q_{1^{l_1}2^{l_2}\ldots n^{l_n}}$ is at once found for given numerical values of the indices $l_1, l_2, \ldots l_n$ by writing down (5) for $l_1 + l_2 + \ldots + l_n$ variables and identifying the values of l_1 of them with that of the first and so on.

Proof of Isserlis formula

$\mathbb{E}(X_1 X_2 \dots X_k) = \sum_{\pi \in \mathcal{P}_2(k)} \prod_{\{i,j\} \in \pi} \mathbb{E}(X_i X_j)$

- Write $k = 2n$ as both sides are zero for odd k. The proof is by induction on n. Note that for $k < 2n$ vector $(X_1, \dots X_k)$ is jointly normal with Σ_k taken as the appropriate submatrix of Σ.
- Case $n = 1$ is obvious $\mathbb{E}(X_1 X_2) = \Sigma_{12}$
- Induction step:

$$\mathbb{E}(X_1 X_2 \dots X_{2n}) = \sum_{j=2}^{2n} \mathbb{E}(X_1 X_j) \mathbb{E} \prod_{i \neq 1, j} X_i$$

  $\pi = \{1, j\} \cup \pi'$

- Then look at $\frac{\partial}{\partial t_1}$ in $\mathbb{E}(X_1 X_2 \dots X_{2n}) = (-1)^n \frac{\partial^{2n}}{\partial t_1 \dots \partial t_{2n}} \varphi(\mathbf{t})\big|_{\mathbf{t}=0}$

$$(-1)^n \frac{\partial^{2n}}{\partial t_1 \dots \partial t_{2n}} \varphi(\mathbf{t})\Big|_{\mathbf{t}=0} = (-1)^n \frac{\partial^{2n-1}}{\partial t_2 \dots \partial t_{2n}} \left(\psi(t_2 \dots, t_{2n}) \frac{\partial}{\partial t_1} (e^{-\frac{\Sigma_{11} t_1^2}{2} - \sum_{j=2}^{2n} \Sigma_{1,j} t_1 t_j})|_{t_1=0} \right) \Big|_{\mathbf{t}=0}$$

$$= (-1)^n \frac{\partial^{2n-1}}{\partial t_2 \dots \partial t_{2n}} \left(-\psi(t_2, \dots, t_{2n}) \sum_{j=2}^{2n} \Sigma_{1,j} t_j \right) |_{\mathbf{t}=0}$$

$$= (-1)^{n-1} \sum_{j=2}^{2n} \Sigma_{1j} \frac{\partial^{2n-2}}{\partial t_2 \cdots \cancel{\partial t_j} \dots \partial t_{2n}} \frac{\partial}{\partial t_j} \left(t_j \psi(t_2 \dots, t_{2n}) \right) |_{t_j=0} \Big|_{\mathbf{t}=0}$$

$$= (-1)^{n-1} \sum_{j=2}^{2n} \Sigma_{1j} \frac{\partial^{2n-2}}{\partial t_2 \cdots \cancel{\partial t_j} \dots \partial t_{2n}} (\varphi(\mathbf{t})) \Big|_{\mathbf{t}=0}$$

$$= \sum_{j=2}^{2n} E(X_1 X_j) \mathbb{E}(X_2 \dots \check{X}_j \dots X_{2n})$$

Wigner matrices

A Wigner matrix is a **symmetric** random matrix $\mathbf{W} = \frac{1}{\sqrt{n}} \begin{bmatrix} Z_{11} & Z_{12} & \dots & Z_{1n} \\ Z_{12} & Z_{22} & \dots & Z_{2n} \\ \vdots & & & \vdots \\ Z_{1n} & Z_{2n} & \dots & Z_{nn} \end{bmatrix}$ where Z_{ij} are independent $N(0, 1)$ random variables.

Clearly, $\mathbf{W} = \sum_{i \leq j \leq n} Z_{ij} E_{ij}$ with deterministic matrices E_{ij}.

It turns out that the following holds:

$$\lim_{n \to \infty} \frac{1}{n} \mathrm{tr}(\mathbf{W}^{\mathrm{k}}) = \int_{-2}^{2} \mathrm{x}^{\mathrm{k}} \frac{\sqrt{4 - \mathrm{x}^2}}{\pi} \mathrm{dx} \text{ in probability, in } L_1\text{, and almost surely for an infinite array } Z_{ij}$$

Wigner was interested in the eigenvalues $\Lambda_1, \dots, \Lambda_n$ of $\mathbf{X}$ and empirical spectral distribution $F_n(x) = \frac{1}{n} \#\{\Lambda_k \leq x\}$. The above shows that (random) moments $\int x^k dF_n$ converge. One can show that this implies $F_n \xrightarrow{\mathcal{D}} \frac{\sqrt{4-x^2}}{\pi} dx$ with probability one. The measure $\frac{\sqrt{4-x^2}}{\pi} dx$ is called Wigner's semicircle law and plays a role of the standard normal distribution in free probability.

Gaussian random matrices

Consider the set $\mathbb{M} \equiv \mathbb{R}^{n(n+1)/2}$ of all symmetric $n \times n$ matrices with inner product $\langle A, B \rangle = tr(AB)$. (Does the definition of normal distribution depend on the inner product?)

$\langle A, B \rangle = \sum_{i=1}^n \sum_{j=1}^n a_{ij} b_{ji} = \sum_{i=1}^n a_{ii} b_{ii} + 2 \sum_{i<j} a_{ij} b_{ij}$

Definition 42. $\mathbf{X}$ is (centered) normal matrix-valued random variable if $\mathbf{X} = \sum_j Z_j A_j$ for some deterministic symmetric matrices A_j.

The characteristic function of $\mathbf{X}$ is $\varphi(T) = \mathbb{E} e^{i \mathrm{tr}(T\mathbf{X})}$. So $\boxed{\varphi(T) = \exp -\frac{1}{2} \sum_j \mathrm{tr}^2(\mathrm{TA_j})}$ In particular, we may ask about $\varphi(T) = e^{-\frac{1}{2} \mathrm{tr}(\mathrm{T}^2)}$. Because $E_{i,j}$ are an orthogonal basis of $\mathbb{M}$, we can expand

$$T = \sum_{i=1}^{n} \mathrm{tr}(\mathrm{TE_{ii}}) \mathrm{E_{ii}} + \sum_{\mathrm{i<j}} \frac{\mathrm{tr}(\mathrm{TE_{ij}})}{\mathrm{tr}(\mathrm{E_{ij}^2})} \mathrm{E_{ij}}$$

$$T=\sum_{i=1}^{n}\mathrm{tr}(\mathrm{TE_{ii}})\mathrm{E_{ii}}+\sum_{\mathrm{i<j}}\frac{\mathrm{tr}(\mathrm{TE_{ij}})}{2}\mathrm{E_{ij}}$$

So $\|T\|^2=\mathrm{tr}(\mathrm{T}^2)=\sum_{\mathrm{i}}\mathrm{tr}^2(\mathrm{TE_{ii}})+\sum_{\mathrm{i<j}}\mathrm{tr}^2(\mathrm{TE_{ij}})/2$ This means that we want $\mathbf{X}=\sum_{i=1}^{n}E_{ii}Z_i+\sum_{i<j}E_{ij}Z_{ij}/\sqrt{2}$

$$\mathbf{x}=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{2}z_{11} & z_{12} & \dots & z_{1n}\\ z_{12} & \sqrt{2}z_{22} & \dots & z_{2n}\\ \vdots & & & \vdots\\ z_{1n} & z_{2n} & \dots & \sqrt{2}z_{nn}\end{bmatrix}$$

Gaussian Orthogonal Ensemble

This is the celebrated Gaussian Orthogonal Ensemble (GOE),

$$\mathbf{x}=\frac{1}{\sqrt{2}}\begin{bmatrix}\sqrt{2}z_{11} & z_{12} & \dots & z_{1n}\\ z_{12} & \sqrt{2}z_{22} & \dots & z_{2n}\\ \vdots & & & \vdots\\ z_{1n} & z_{2n} & \dots & \sqrt{2}z_{nn}\end{bmatrix}$$

which is sometimes confused with the Wigner matrix $\mathbf{W}$ of i.i.d $N(0,1)$ random variables. Up to a scaling, $\mathbf{X}$ and $\mathbf{W}$ differ only by an extra factor on the main diagonal.

GOE matrix $\mathbf{X}$ arises naturally by symmetrization: with non-symmetric i.i.d. matrix $\mathbf{Z}=[Z_{i,j}]$, we take $\mathbf{X}=(\mathbf{Z}+\mathbf{Z}')/2$.

GOE refers to invariance under orthogonal group: $\mathbf{X}\simeq U\mathbf{X}U'$ for orthogonal matrix U. This property is easy to check using characteristics function and "tracial property" $\mathrm{tr(AB)}=\mathrm{tr(BA)}$.

$\varphi_{U\mathbf{X}U'}(T)=\mathbb{E}\exp i\mathrm{tr}(\mathrm{TU}\mathbf{X}\mathrm{U}')=\mathbb{E}\exp i\mathrm{tr}(\mathrm{U'TU}\mathbf{X})=\mathrm{e}^{-\frac{1}{2}\mathrm{tr}((\mathrm{U'TU})^2)}=e^{-\frac{1}{2}\mathrm{tr}(\mathrm{U'T^2U})}=e^{-\frac{1}{2}\mathrm{tr}(\mathrm{UU'T^2})}=e^{-\frac{1}{2}\mathrm{tr}(\mathrm{T}^2)}=\varphi_{\mathbf{X}}(T)$

- GOE matrix $\mathbf{x}\in\mathbb{M}$, has density $f(\mathbf{x})=C\exp(-\frac{1}{2}tr(\mathbf{x}^2))$ with respect to Lebesgue measure on $\mathbb{R}^{n(n+1)/2}$ i.e. with respect to $dx_{11}dx_{12}\dots dx_{1n}dx_{22}dx_{23}\dots dx_{2n}\dots dx_{nn}$.
- Polynomial perturbations $f_\varepsilon(\mathbf{x})=C_\varepsilon\exp(-\frac{1}{2}tr(\mathbf{x}^2)+\varepsilon\mathrm{tr}(\mathbf{x}^4))$ preserve orthogonal invariance at the expense of loosing connection with independence.
- In another direction, one can study random matrices that are constructed from non-normal independent random variables. For example, in population genetics the SNP data consist of $M\times N$ matrices of order $M\sim10^3$ and $N\sim10^6$ with entries that take 3 values $\{0,1,2\}$ and are independent between rows and "weakly linked" between columns.

The end

Final Exam projects are already posted.

Thank you

References

[Billingsley] P. Billingsley, Probability and Measure IIIrd edition

[Durrett] R. Durrett, Probability: Theory and Examples, Edition 4.1 (online)

[Gut] A. Gut, Probability: a graduate course

[Resnik] S. Resnik, A Probability Path, Birkhause 1998

[Proschan-Shaw] S M. Proschan and P. Shaw, Essential of Probability Theory for Statistitcians, CRC Press 2016

[Varadhan] S.R.S. Varadhan, Probability Theory, (online pdf from 2000)