Session 43

Application of WLLN to Relative Freqs.

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

Scenario: Repeat a simple experiment

$(\Omega, \mathcal{F}, P)$

independently many times.

Let: $A \in \mathcal{F}$

Define: $X_k = 1_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \omega \notin A \end{cases}$

on the kth repetition of the experiment

...

As defined above

$X_1 \; X_2 \; \ldots$ is a iid seq.

with $E[X_k] = E[1_A(\omega)] = P(A)$

Relative Freqs. + WLLN (cont'd)

Define

$$r_n(A) = \frac{X_1 + X_2 + \cdots + X_n}{n}$$

$$= \frac{1}{n} \sum_{i=1}^{n} X_i$$

this is the relative freq. with which A occurs in n trials.

$$\Rightarrow E\left[r_n(A)\right] = E\left[\frac{X_1 + \cdots + X_n}{n}\right]$$

$$= \frac{P(A) + \cdots + P(A)}{n}$$

$$= P(A)$$

Can also show: $\mathrm{Var}[X_k] = P(A)\left[1 - P(A)\right] < \infty$

$\therefore$ WLLN applies

$$P\left(\left|r_n(A) - P(A)\right| \geq \varepsilon\right) \longrightarrow 0 \quad \text{as } n \to \infty$$

$\forall \varepsilon > 0$.

ie relative freq. $\longrightarrow$ actual prob.

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

The Central Limit Theorem (CLT)

CLT is a (family of) theorem(s) that says, under very general conditions, the sum of a large number of RVs is a Gaussian distribution.

regardless of their distribution

CLT Theorem (one form)

$X_1 \; X_2 \; \cdots \; X_n \; \cdots$ sequence of iid RVs, each having mean μ and variance $\sigma^2 < \infty$.

Then

$$Z_n \triangleq \frac{(X_1 + \cdots + X_n) - n\mu}{\sigma\sqrt{n}}$$

is approximately Gaussian with mean 0 and variance 1. That is

$$\lim_{n \to \infty} P(Z_n \le z) = \Phi(z) \triangleq \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \, dx$$

Convergence in distribution.

Lemma Used in Proof of CLT

① $Z_1, Z_2, \ldots$ be a sequence of rvs each having

cumulative distribution functs $\quad F_{Z_n}(z_n) \triangleq P(Z_n \leq z_n)$

moment generating functs $\quad \phi_{Z_n}(s) \triangleq E\left[e^{sZ_n}\right]$

② Z a rv with cdf $F_Z(z)$ and mgf $\phi_Z(s)$

Then, if

$$\lim_{n\to\infty} \phi_{Z_n}(s) = \phi_Z(s) \quad \forall s$$

it follows that

$$\lim_{n\to\infty} F_{Z_n}(z) = F_Z(z)$$

for all $z \in \mathbb{R}$ st. $F_Z(z)$ is continuous.

(We won't try to prove this)

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

Sketch of Proof of CLT

Assume $\mu = 0$, $\sigma^2 = 1$

Assume the required mgfs all exist.

$\{X_i\}$ iid with a common cdf corresp to a rv X

$$\phi_X(s) = E\left[e^{sX}\right]$$

CLT statistic in CLT $\quad Z_n = \dfrac{X_1 + \dots + X_n}{\sqrt{n}}$

$$= \frac{X_1}{\sqrt{n}} + \dots + \frac{X_n}{\sqrt{n}}$$

$$\phi_{X/\sqrt{n}}(s) = E\left[e^{sX/\sqrt{n}}\right]$$

$$= \phi_X(s/\sqrt{n})$$

$\therefore$ By iid

$$\phi_{Z_n}(s) = \left[\phi_{X/\sqrt{n}}(s)\right]^n$$

$$= \left[\phi_X(s/\sqrt{n})\right]^n$$

Define $\quad L(s) = \log \phi_X(s)$

$\Rightarrow$ <u>Note</u> $\quad L(0) = \log \phi_X(0) = 0$

$$L'(0) = \frac{\phi_X'(0)}{\phi_X(0)} = \frac{\mu}{1} = 0$$

$L''(0) = E\left[X^2\right] = 1$

$\hookrightarrow$ do the deriv. & eval @ $s=0$.

Need to show

$$\phi_{Z_n}(s) = \left[\phi_X(s/\sqrt{n})\right]^n \longrightarrow e^{s^2/2} \quad \text{as } n\to\infty$$

$\forall s.$

$$\log \phi_{Z_n}(s) \longrightarrow s^2/2 \quad \text{as } n\to\infty.$$

$$\shortparallel$$

$$n \log \phi_X(s/\sqrt{n}) = n\,L(s/\sqrt{n}). = \frac{L(s/\sqrt{n})}{n^{-1}}$$

Gives indeterminant form ... L'ôpital Rule.

We will need to use LH rule twice ...

Will see that

$$\lim_{n\to\infty} \phi_{Z_n}(s) = \lim_{n\to\infty} \frac{1}{2} s^2 L''(s/\sqrt{n}).$$

$$= \frac{1}{2} s^2 \quad \text{in the mgf of } N(0,1).$$

Ross Proof of CLT

To start assume $\mu = 0$ and $\sigma^2 = 1$. We will also assume that the mgfs of the X_i exist. Since they are iid, they all have the same mgf

$$\phi_X(s) = E\left[e^{sX}\right]$$

X a rv with the common dist of the X_i

In this case the statistic from the statement of CLT would be

$$Z_n = \frac{X_1 + X_2 + \cdots + X_n}{\sqrt{n}}$$

Clearly
$$\phi_{X/\sqrt{n}} = E\left[e^{sX/\sqrt{n}}\right]$$

$$= \phi_X\left(s/\sqrt{n}\right)$$

Therefore, the mgf of Z_n would be

$$\phi_{Z_n}(s) = \left[\phi_X(s/\sqrt{n})\right]^n$$

Define

$$L(s) = \log \phi_X(s)$$

and note that

$$L(0) = \log \phi_X(0) = 0$$

$$L'(0) = \frac{\phi_X'(0)}{\phi_X(0)} = \frac{\mu}{1} = \mu = 0$$

$$L''(0) = \frac{\phi_X(0)\,\phi_X''(0) - \left[\phi_X'(0)\right]^2}{\left[\phi_X(0)\right]^2}$$

$$= E[X^2] = 1$$

$\boxed{\text{Re: } L''(s)}$ $L'(s) = \dfrac{\phi_X'(s)}{\phi_X(s)}$

Taking another derivative wrt s ...

$$\frac{\phi_X''(s)\,\phi_X(s) - \phi_X'(s)\,\phi_X'(s)}{\left[\phi_X(s)\right]^2}$$

Eval @ s=0 to get formula.

Then to prove the CLT under the above assumptions we need to ... show

$$\phi_{Z_n}(s) = \left[\phi_X(s/\sqrt{n})\right]^n \longrightarrow e^{s^2/2}$$

as $n \to \infty$ (for all s).

This is enough to show

$$\log \phi_{Z_n}(s) \longrightarrow \frac{s^2}{2}$$

$$\log \phi_{Z_n}(s) = n \log \phi_X(s/\sqrt{n})$$

$$= n L(s/\sqrt{n})$$

Now to consider

$$\lim_{n\to\infty} n L(s/\sqrt{n}) = \lim_{n\to\infty} \frac{L(s/\sqrt{n})}{n^{-1}}$$

which gives an indeterminent form as $n \to \infty$

$\Rightarrow$ Try L'Hopital rule

$$\frac{d}{dn} L(s/\sqrt{n}) = L'(s/\sqrt{n})\left(-\tfrac{1}{2} n^{-3/2} s\right)$$

$$= -\tfrac{1}{2} L'(s/\sqrt{n})\, n^{-3/2} s$$

$$\frac{d}{dn} n^{-1} = -n^{-2}$$

Check

$$\lim_{n\to\infty} \frac{-\frac{1}{2} L'(s/\sqrt{n})\, n^{-3/2} s}{-n^{-2}}$$

still have indeterminent form

$$\frac{L'(s/\sqrt{n})\, s}{2 n^{-1/2}}$$

L'Hopital Again: $\dfrac{L'(s/\sqrt{n})\, n^{-3/2}\, s}{2n^{-2}}$

$$= \frac{L'(s/\sqrt{n})\, s}{2n^{-1/2}}$$

(again indeterminent $0/\infty$ form as $n \to \infty$)

$$\frac{d}{dn} L'(s/\sqrt{n})\, s = L''(s/\sqrt{n})\left(-\tfrac{1}{2} n^{-3/2} s\right) s$$

$$= -\frac{1}{2} s^2 L''(s/\sqrt{n})\, n^{-3/2}$$

$$\frac{d}{dn} 2n^{-1/2} = -2\left(\tfrac{1}{2}\right) n^{-3/2}$$

$\Rightarrow$ ratio to take lim $n \to \infty$ is

$$\frac{-\frac{1}{2} s^2 L''(s/\sqrt{n})\, n^{-3/2}}{-n^{-3/2}} = \frac{1}{2} s^2 L''(s/\sqrt{n})$$

But as $n \to \infty$ this conv. to $s^2/2$

$\square$

Transformation for general case

$$E[X_i] = \mu \neq 0$$

$$Var[X_i] = \sigma^2 \neq 1$$

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

Define

$$\tilde{X}_i = \frac{X_i - \mu}{\sigma}$$

which do satisfy the conditions under which CLT was proven.

$$\tilde{Z}_n = \frac{\tilde{X}_1 + \cdots + \tilde{X}_n}{\sqrt{n}} = \frac{X_1 + \cdots + X_n - n\mu}{\sigma\sqrt{n}}$$

$\Rightarrow$ general case holds also.

Interpretation ....

$$X_1 + X_2 + \cdots + X_n \sim N(n\mu, n\sigma^2).$$

Exercise 5. (FROM TEXTBOOK, PROBLEM 7.23)
(NOTE: This is a different problem than above.) A large number n of voters are polled and the proportion in favor is estimated using a relative frequency. That is to say, the fraction of the n voters polled who are in favor is the estimate for the overall population. Use equation the Central Limit Theorem to determine how many voters should be polled so that the probability is at least 0.95 that $f_A(n)$ differs from 0.20 by less than 0.02.

Okay its a different problem, but the setup is the same. --- look @ hint for Exercise 4 for setup.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) \qquad E\left[1_A(X_i)\right] = p$$

$$\mathrm{Var}\left[1_A(X_i)\right] = p(1-p)$$

Find n st.

$$P\left(0.18 < \hat{p} \le 0.22\right) \geq 0.95$$

To fit the CLT framework we need to re-normalize

For short let $Z_i = 1_A(X_i)$ and write the event of interest as ...

$$\left\{0.18 < \hat{p} \le 0.22\right\} = \left\{0.18 < \frac{1}{n}\sum_{i=1}^{n} Z_i \le 0.22\right\}$$

$$= \left\{0.18n < \sum_{i=1}^{n} Z_i \le 0.22n\right\}$$

$$= \left\{0.18n - np < \sum_{i=1}^{n} Z_i - np \le 0.22n - np\right\}$$

$$= \left\{ \frac{0.18n - np}{\sqrt{(1-p)pn}} < \frac{\sum_{i=1}^{n} Z_i - np}{\sqrt{(1-p)pn}} \le \frac{0.22n - np}{\sqrt{(1-p)pn}} \right\}$$

$$= \left\{ \sqrt{n}\,\underbrace{\frac{(0.18-0.2)}{0.4}}_{-.02/0.4} < \underbrace{\frac{\sum_{i=1}^{n} Z_i - np}{\sqrt{(1-p)pn}}}_{\sim N(0,1)} \le \sqrt{n}\,\underbrace{\frac{(0.22-0.2)}{0.4}}_{+0.02/0.4} \right\}$$

negatives of each other

$p = 0.2$

$p(1-p) = (0.2)(0.8)$

$= 0.16$

$\sqrt{p(1-p)} = \cancel{0.04}$ 0.4

$$P(-\alpha < N \le \alpha) = \Phi(\alpha) - \Phi(-\alpha)$$

$$= \Phi(\alpha) - [1 - \Phi(\alpha)]$$

$$= 2\Phi(\alpha) - 1 = 0.95$$

$$\Rightarrow 2\Phi(\alpha) = 1.95$$

$$\Phi(\alpha) = 1.95/2 = 0.975$$

$$\alpha = \Phi^{-1}(0.975) = 1.96$$

$$\therefore \sqrt{n}\,\frac{0.02}{0.4} = 1.96 \Rightarrow n = \left(\frac{0.4}{0.02}\,1.96\right)^2$$

$$= 20^2 \cdot 1.96^2 = 1536.6$$

1537

Exercise 5. (From textbook, problem 7.23)
(NOTE: This is a different problem than above.) A large number n of voters are polled and the proportion in favor is estimated using a relative frequency. That is to say, the fraction of the n voters polled who are in favor is the estimate for the overall population. Use equation the Central Limit Theorem to determine how many voters should be polled so that the probability is at least 0.95 that $f_A(n)$ differs from 0.20 by less than 0.02.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} 1_A(X_i) \qquad E[1_A(X_i)] = p$$
$$Var[1_A(X_i)] = p(1-p)$$

Find value n s.t.

$$P(0.18 < \hat{p} \leq 0.22) \geq 0.95 \qquad p = 0.2$$

Fit CLT framework $Z_i \triangleq 1_A(X_i)$ and fiddle with the event of interest ...

$$\{0.18 < \hat{p} \leq 0.22\} = \left\{0.18 < \frac{1}{n}\sum_{i=1}^{n} Z_i \leq .22\right\}$$

$$= \left\{0.18n < \sum_{i=1}^{n} Z_i \leq 0.22n\right\}$$

$$= \left\{\underbrace{\frac{0.18n - np}{\sqrt{(1-p)pn}}}_{-\alpha} < \frac{\sum_{i=1}^{n} Z_i - np}{\sqrt{(1-p)pn}} \leq \overbrace{\frac{0.22n - np}{\sqrt{(1-p)pn}}}^{\alpha}\right\}$$

$\hookrightarrow$ for large n $\sim N(0,1)$.

$$P(-\alpha < N \leq \alpha) = \Phi(\alpha) - \Phi(-\alpha).$$
$$= \Phi(\alpha) - [1 - \Phi(\alpha)] = 2\Phi(\alpha) - 1 \overset{\text{set}}{=} 0.95$$

$$\Leftarrow \quad 2\Phi(\alpha) = 1.95$$
$$\Phi(\alpha) = .975 \longrightarrow \alpha = \Phi^{-1}(.975).$$

$\alpha = 1.96.$

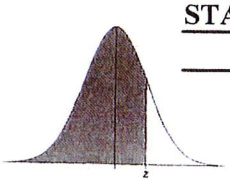
$$\frac{0.22n - np}{\sqrt{(1-p)p\,n}} = \sqrt{n}\left(\frac{0.02}{0.4}\right) \overset{\text{set}}{=} 1.96$$

$$n = 20^2 (1.96)^2 = 1536.6$$

1537

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY

1.96.



STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.50000	.50399	.50798	.51197	.51595	.51994	.52392	.52790	.53188	.53586
0.1	.53983	.54380	.54776	.55172	.55567	.55962	.56356	.56749	.57142	.57535
0.2	.57926	.58317	.58706	.59095	.59483	.59871	.60257	.60642	.61026	.61409
0.3	.61791	.62172	.62552	.62930	.63307	.63683	.64058	.64431	.64803	.65173
0.4	.65542	.65910	.66276	.66640	.67003	.67364	.67724	.68082	.68439	.68793
0.5	.69146	.69497	.69847	.70194	.70540	.70884	.71226	.71566	.71904	.72240
0.6	.72575	.72907	.73237	.73565	.73891	.74215	.74537	.74857	.75175	.75490
0.7	.75804	.76115	.76424	.76730	.77035	.77337	.77637	.77935	.78230	.78524
0.8	.78814	.79103	.79389	.79673	.79955	.80234	.80511	.80785	.81057	.81327
0.9	.81594	.81859	.82121	.82381	.82639	.82894	.83147	.83398	.83646	.83891
1.0	.84134	.84375	.84614	.84849	.85083	.85314	.85543	.85769	.85993	.86214
1.1	.86433	.86650	.86864	.87076	.87286	.87493	.87698	.87900	.88100	.88298
1.2	.88493	.88686	.88877	.89065	.89251	.89435	.89617	.89796	.89973	.90147
1.3	.90320	.90490	.90658	.90824	.90988	.91149	.91309	.91466	.91621	.91774
1.4	.91924	.92073	.92220	.92364	.92507	.92647	.92785	.92922	.93056	.93189
1.5	.93319	.93448	.93574	.93699	.93822	.93943	.94062	.94179	.94295	.94408
1.6	.94520	.94630	.94738	.94845	.94950	.95053	.95154	.95254	.95352	.95449
1.7	.95543	.95637	.95728	.95818	.95907	.95994	.96080	.96164	.96246	.96327
1.8	.96407	.96485	.96562	.96638	.96712	.96784	.96856	.96926	.96995	.97062
1.9	.97128	.97193	.97257	.97320	.97381	.97441	.97500	.97558	.97615	.97670
2.0	.97725	.97778	.97831	.97882	.97932	.97982	.98030	.98077	.98124	.98169
2.1	.98214	.98257	.98300	.98341	.98382	.98422	.98461	.98500	.98537	.98574
2.2	.98610	.98645	.98679	.98713	.98745	.98778	.98809	.98840	.98870	.98899
2.3	.98928	.98956	.98983	.99010	.99036	.99061	.99086	.99111	.99134	.99158
2.4	.99180	.99202	.99224	.99245	.99266	.99286	.99305	.99324	.99343	.99361
2.5	.99379	.99396	.99413	.99430	.99446	.99461	.99477	.99492	.99506	.99520
2.6	.99534	.99547	.99560	.99573	.99585	.99598	.99609	.99621	.99632	.99643
2.7	.99653	.99664	.99674	.99683	.99693	.99702	.99711	.99720	.99728	.99736
2.8	.99744	.99752	.99760	.99767	.99774	.99781	.99788	.99795	.99801	.99807
2.9	.99813	.99819	.99825	.99831	.99836	.99841	.99846	.99851	.99856	.99861
3.0	.99865	.99869	.99874	.99878	.99882	.99886	.99889	.99893	.99896	.99900
3.1	.99903	.99906	.99910	.99913	.99916	.99918	.99921	.99924	.99926	.99929
3.2	.99931	.99934	.99936	.99938	.99940	.99942	.99944	.99946	.99948	.99950
3.3	.99952	.99953	.99955	.99957	.99958	.99960	.99961	.99962	.99964	.99965
3.4	.99966	.99968	.99969	.99970	.99971	.99972	.99973	.99974	.99975	.99976
3.5	.99977	.99978	.99978	.99979	.99980	.99981	.99981	.99982	.99983	.99983
3.6	.99984	.99985	.99985	.99986	.99986	.99987	.99987	.99988	.99988	.99989
3.7	.99989	.99990	.99990	.99990	.99991	.99991	.99992	.99992	.99992	.99992
3.8	.99993	.99993	.99993	.99994	.99994	.99994	.99994	.99995	.99995	.99995
3.9	.99995	.99995	.99996	.99996	.99996	.99996	.99996	.99996	.99997	.99997

$(\Omega, \mathcal{F}, P)$

A random process is a collection of rvs def. on $(\Omega, \mathcal{F}, P)$ indexed by a parameter $t \in \mathbb{R}$.

$$X(t, \omega) \quad \begin{array}{l} t \in \mathbb{R} \\ \omega \in \Omega \end{array}$$

FORM C
APPROVED FOR USE IN
PURDUE UNIVERSITY