Bios 601 H3 V1

Jiewen Liu 260825295

September 25, 2021

0.1

0.1.1

It has a pyramid shape. The shape is very wide at the younger ages. The critical point takes around 1881, after which the younger ages percentage changes seem to be explained by some natural factors. Before 1881, the shape of age distribution may be accredited to the Great Famine of the mid-19th century, and after it, the darkest time of the Great Famine, we see the population starts to show and keep the trend of expansion till the low ebb at the 1879's potato blight. There are surges in 1851, 1871 and 1881 (some errors introduced?).

0.1.2

Mean: 31; SD: 12 ; Q75: 13; Q50: 25; Q25: 45.

0.1.3

I think it is CLT-friendly in this case. The sample size $n > 30$ is could be good enough to give a r.v., the sample mean, that is almost from a normal distribution. I also see the **concern** behind. There exist cases that are extreme enough, for example heavily skewed, to render the sample mean deviate much from Gaussian distribution with n at the hundred (n=k*100) or thousand (n=k*1000) level. A easy practice to construct such an extreme distribution is taking a non-Gaussian infinitely divisible distribution and decomposing it. Another quick instance is binomial distribution. It is quite intuitive that you need to increase sample size n when p is quite low since the sample mean definitely converges much more slowly as compared to $p = 0.5$

0.6

0.6.1

Here I assume the SD in the question refers to one of the population, not the sample mean. The standard deviation is a measure of the amount of variation of a set of values, which has already been an averaged (or weighted by probability at each point) value/noti. The true variation within the dataset has nothing to do with the size n.

0.6.2

I suspect the SD given in the column, in fact, are the calculated results for SEM. Thus we could times them by timing with $\sqrt{n}$ to recover the estimate standard error (SE) for standard deviation (SD) of the sample mean.

0.6.3

The mathematician is correct. The reason is that the two random variables $A = sd(rnorm(10))$ and $B = sd(rnorm(10))$ are not from the exact same distribution and they are even biased estimators for the population's SD. For an unbiased estimator, if you take the square root afterwards, it could be biased. In addition, even if we give back the square, say calculate $\hat{SD}^2$, though unbiasdness, the distribution's parameter is not the same!

Thus there is no guarantee that 50% probability that $A > B$.

How could 53% be analytically verified? If we divide $(n-1)\hat{SD}^2$ with (n-1), it is not Chi-squared distribution but scaled Gamma one. Think about the distribution behind and do some integrals :)

$$\begin{aligned} P(A \leq B) &= P(A^2 \leq B^2) && (1)\\ &= P(Gamma((n_1-1)/2, 2/(n_1-1)) \leq Gamma((n_2-1)/2, 2/(n_2-1))) && (2)\\ &= P(Gamma((9/2, 2/9) \leq Gamma(19/2, 2/19)) && (3)\\ &\approx 0.528 && (4) \end{aligned}$$

Verification by simulation using R (rounding to 2 is not necessary):

```
a= rep(0,100000)
b= rep(0,100000)
for(i in seq(1,100000,1)){
  a[i]=sd( rnorm(10) )
  b[i]=sd( rnorm(20) )
}
sum(a>b)/100000 -> Result: 0.47289.
```

0.7

0.7.1

The 344 and 351 must be picked. The remaining one 377 could be from the either day.

$$\frac{\binom{2}{2} * \binom{2}{1} * \binom{21}{0}}{\binom{25}{3}} \approx 0.001$$

It just seems to be an interesting pattern. Really something unusual causes this? Need to be further examined.

0.7.2

Here are some general ideas. To avoid p-hacking, i.e. data dredging, we definitely need more considerations. We could compare the data on a week-on-week / month-on-month / year-on-year basis while carefully considering any confounding factors. In addition, we could have a ANOVA or, if some other random variables collected, decode weekend/Saturday/Sunday, "BlackoutAffected?", etc. into 1 or 0 factor and do the regression to see if the key factor "BlackoutAffected?" is significant or not. To have causal thinking in, talk to experts/doctors about any other possible reasons/variables that may naturally render such pattern.

0.7.3

Collect and record potentially useful random variable including day of the week + month (considering periodical and seasonal fluctuations), factors about population structure, "BlackoutAffected?", etc. The data should be not only cross-sectional but also longitudinal (if possible, throughout ayear before and a year after in different regions'/cities' hospitals). Carry out DID (Difference in Difference) analysis.

0.8

Assume if 10 broke tires allowed at max during the 7500km travel. Since the memoryless property is assumed in the question, it is a poisson process and the *exponential*(λ) is reasonable choice to model the time for the duration time. 4 tires are on the risk at the same time. The rate for poisson is fourfold.

0.8.1

Set $\lambda = 1 * 4$ and $t = 1.5$, plug in and obtain $f_X = \frac{6^x e^{-6}}{x!}$ where $x \in \{0, 1, 2, 3 \ldots\}$.

$P(\text{Successful Travelling}) = P(X \leq 10) = \sum_{x=0}^{10} f_X(x) = 0.916076 \approx 0.92$

Beautiful!

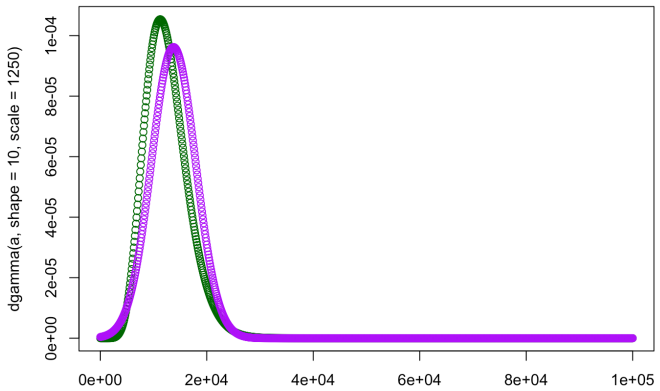
0.8.2

The sum of variables from exponential distributions with the same λ has the $Erlang(k, \lambda) = Gamma(k, \lambda^{-1})$ distribution. Set $\lambda = 4/5000 = 1/1250$. Let $Y = \sum_{i=1}^{10} X_i$. Then $Y \sim Gamma(11, 1250)$ and $f_Y(y) = \frac{1}{\Gamma(11)1250^{10}} y^{10} e^{-y/1250}$.

$P(\text{Successful Travelling}) = P(Y > 7500) = \int_{7500}^{\infty} \frac{1}{\Gamma(10)1250^{10}} t^9 e^{-t/1250} dt = 0.916076 \approx 0.92$

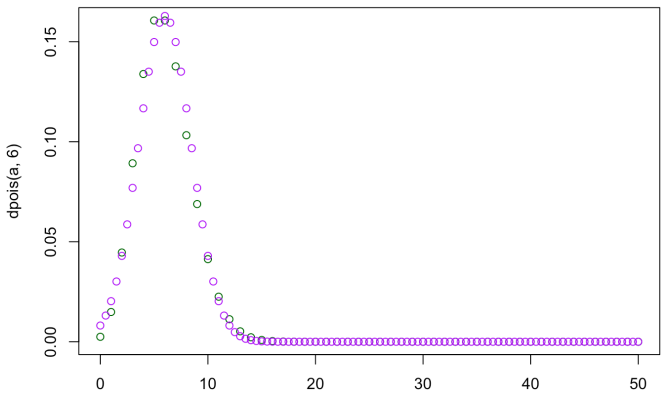
0.8.3

$Normal(11 * 1250, \sqrt{11 * 1250^2})$ gives an acceptable approximation to $Gamma(11, 1250)$ here. Though some skewness, even better when 11 becomes larger and 1250 becomes lower. Purple points are from a normal distribution. Green from Gamma one.



0.8.4

$Normal(6, \sqrt{6})$ gives an good approximation. Purple points are from a normal distribution. Green from Poisson one.



0.8.5

```
rexp(100,1/1250),
[1] 1145.5452500  960.7640946 1581.6255938 5849.9155051  144.9102955 1034.0927642
  [7]  399.5684249  241.4792062 3428.8833312  879.3574862  803.6868036   55.9054175
 [13] 2370.0355993  283.0216580  517.5008846  238.9979246 2600.6991950 1203.5695833
 [19] 2689.0530845 4356.1706846    0.1618103  499.0913777  877.5097755  231.0034144
 [25]  268.0940311   81.5233772 1663.7728852  642.1920715 1004.4122079 1349.1648284
 [31] 1323.1231970  219.7254406   18.1993912  716.8770459 1099.1334973 1420.1177762
 [37]  166.6439679 1099.6332589 3516.4707966  363.0645276  152.4665749  503.0377780
```

```
 [43] 1580.3912759  723.1305825   62.6698224 1043.9033945 1937.4943174  507.9320225
 [49] 4302.4918901 3399.9557070 1412.2050333 1477.1531254  162.8239633 1030.4970445
```

0.10

0.10.1

Assume CLT applies to this case (n=100 is large enough to deal with distribution that is skewed to some extent). The question is equivalent to how likely the "random" sample mean will be less than 155 pounds. The SD for a random group of 100 people is $25/\sqrt{100} = 2.5$. The $155 - 150 = 5 = 2 * 2.5$. For one-side 2 unit deviation from the mean, the closest probability from the given choices is 2%.

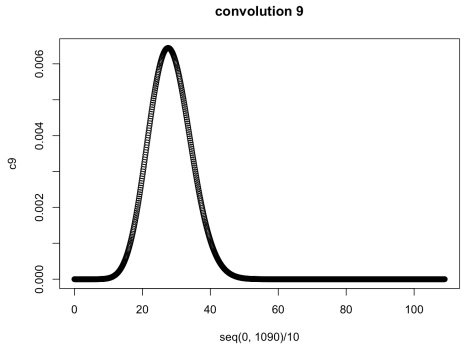
0.10.2

Yes, still comfortable. Though it is unlikely being gaussian, considering the reasonable range for weight of people who is healthy enough to walk in the convention centre and go into the elevator, the distribution of conventioneers' weights should not be crazily extreme. Given the fair large number 100, we should believe the CLT works here. Being "random" is the key because the probability talked about in (i) is based on the background knowledge about conventioneer population provided in the context. Nothing about of probability could be deduced without proper setup. You could definitely talk about other cases, i.e. sub-groups, the heavier or lighter conventioneer groups, get a "random" group of 100 conventioneers from a specific sub-group and make inference only on this group. If "randomness" is lost, all previous inferences will be baseless. For example, you could construct a group of 100 conventioneers from only overweighted ones, and the likelihood for overloading could be 100%.

0.13

0.13.1

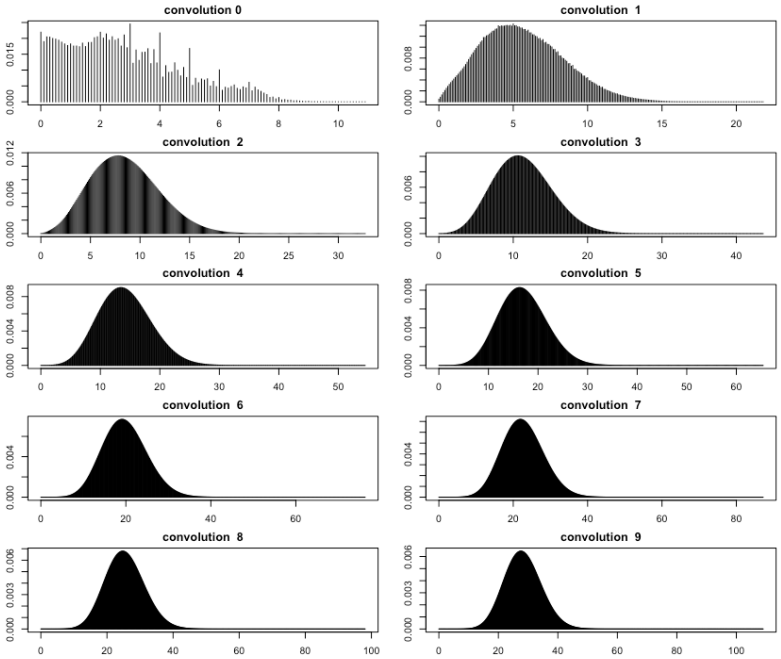
```
path = "http://www.biostat.mcgill.ca/hanley/bios601/AgeFrequencies.txt"
age.distrn = read.table(path,header=TRUE)
x = age.distrn$Age
freq = age.distrn$Freq/sum(age.distrn$Freq)
M = outer(freq,freq,"*")
c9 = freq
for(i in seq(1:9)){
  M = outer(freq,c9,"*")
  c9 = sapply(split(M, col(M) + row(M)), sum)}
plot(seq(0,1090)/10,c9,main='convolution 9')
```



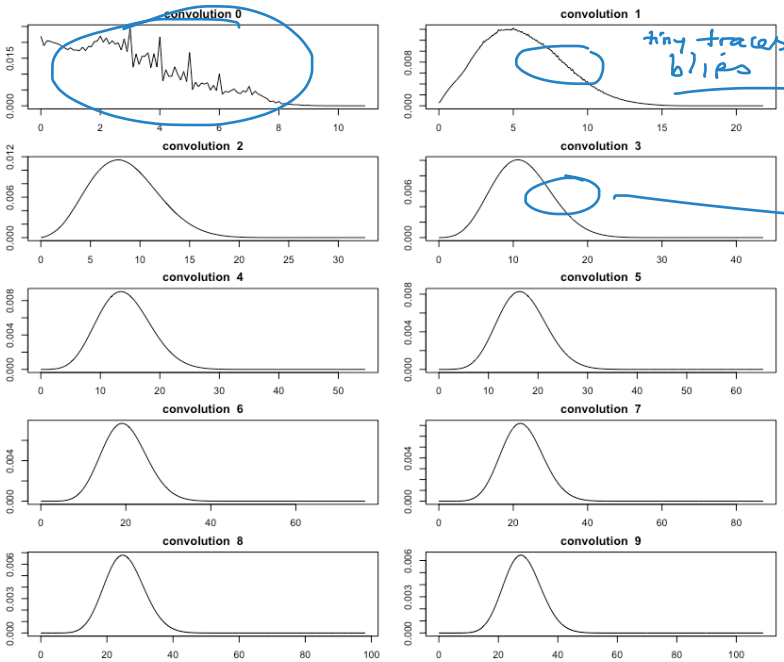
0.13.2

```
par(mfrow=c(5,2),mar=c(2,2,2,2))
c = freq
```

```
plot(seq(0,109)/10,c,type="h",main='convolution 0')
for(i in seq(1:9)){
  M = outer(freq,c,"*")
  c = sapply(split(M, col(M) + row(M)), sum)
  plot(seq(0,109*(i+1))/10,c,type="h",main=paste('convolution ',i))
}
```



0.13.3



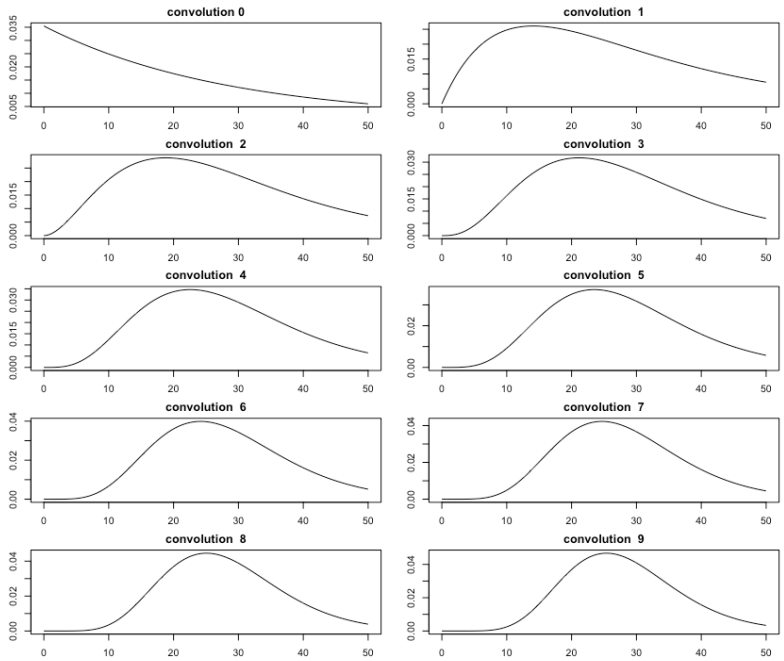
0.13.4

They look really really normal since n=2. Let's take a look

I plug in the inverse sample mean as the parameter λ of our exponential distribution. Then $\lambda = 1/28.2$.

For the sum of exponential distribution, it is a bit skewed. Seems hard to be very nicely approximate it to normal without large enough n. Why? Maybe due to infinite divisibility? We see for the shape of $Gamma(k, \lambda^{-1})$ distribution, the skew is controlled by the shape parameter k, which is the sample size n in our case. K is related to skewness defined to be the 3rd centralized moment in a reverse square root way $(1/\sqrt{k})$. Let the sample size go up. We still end up with the sample mean that has Gamma distribution, which is somewhat right skewed. Due to the square root symbol in the denominator, it costs more to "dilute" stuffs with small size n.

```
sum(age.distrn$Age*freq)
28.2
par(mfrow=c(5,2),mar=c(2,2,2,2))
plot(seq(0,50,0.1),dgamma(seq(0,50,0.1),shape=1,scale=28.2),main='convolution 0',type='l')
for(i in seq(1:9)){
  plot(seq(0,50,0.1),dgamma(seq(0,50,0.1),shape=(i+1),scale=28.2/(i+1)),main=paste('convolution ',i),type=
}
```



0.13.5

Not conspicuous starting from n=2.

I expect your calculations are correct, but I will push on Tuesday for every one to MAKE DIAGRAMS like I do on p15 of Notes

0.16

0.16.1

0.16.1.1

When $\mu = 25$ in fact:

$$\begin{aligned} P(\text{Type I Error}) &= P(\bar{X} > 26) && (5)\\ &= 1 - P(\bar{X} \leq 26) && (6)\\ &= 1 - P(\frac{\bar{X} - 25}{50/\sqrt{900}} \leq \frac{26 - 25}{50/\sqrt{900}}) && (7)\\ &= 1 - P(Z \leq 0.6) && (8)\\ &\approx 27.42\% && (9) \end{aligned}$$

0.16.1.2

When $\mu = 28$ in fact:

$$\begin{aligned} P(\text{Type II Error}) &= P(\bar{X} \leq 26) && (10)\\ &= P(\frac{\bar{X} - 28}{50/\sqrt{900}} \leq \frac{26 - 28}{50/\sqrt{900}}) && (11)\\ &= P(Z \leq -1.2) && (12)\\ &\approx 11.51\% && (13) \end{aligned}$$

0.16.1.3

When $\mu = 30$ in fact:

$$\begin{aligned} P(\text{Type II Error}) &= P(\bar{X} \leq 26) && (14)\\ &= P(\frac{\bar{X} - 30}{50/\sqrt{900}} \leq \frac{26 - 30}{50/\sqrt{900}}) && (15)\\ &= P(Z \leq -2.4) && (16)\\ &\approx 0.82\% && (17) \end{aligned}$$

0.16.1.4

✓

CLT should apply here. 900 is a large sample size. Assume individual customer's expenditure is capped (no crazy wholxesale involved). It is obvious that the lower bound is 0. The shape of the distribution behind the observed value should not be extreme enough to let the magic of CLT go.

0.16.2

The threshold to reject the null hypothesis is that $300 - 1.645 * 3/\sqrt{6} = 297.9853$

0.16.2.1

When the alternative $\mu = 298$ in fact:

$$\begin{aligned} Power(298) &= P(\bar{X} \leq 297.9853) && (18)\\ &= P(\frac{\bar{X} - 297.9853}{3/\sqrt{6}} \leq \frac{297.9853 - 298}{3/\sqrt{6}}) && (19)\\ &\approx 49.52\% && (20) \end{aligned}$$

0.16.2.2

When the alternative $\mu = 294$ in fact:

$$Power(294) = P(\bar{X} \leq 297.9853) \tag{21}$$

$$= P(\frac{\bar{X} - 297.9853}{3/\sqrt{6}} \leq \frac{297.9853 - 294}{3/\sqrt{6}}) \tag{22}$$

$$\approx 99.94\% \tag{23}$$

0.16.2.3

When the alternative $\mu = 296$:

$$Power(294) = P(\bar{X} \leq 297.9853) \tag{24}$$

$$= P(\frac{\bar{X} - 297.9853}{3/\sqrt{6}} \leq \frac{297.9853 - 296}{3/\sqrt{6}}) \tag{25}$$

$$\approx 94.75\% \tag{26}$$

Again I feel less secure checking algebra

Lower. It is closer to 300 as compared to 294. Definitely more easily for sample mean to fall into the accepting region for the null hypothesis than usual answers —

0.16.2.4

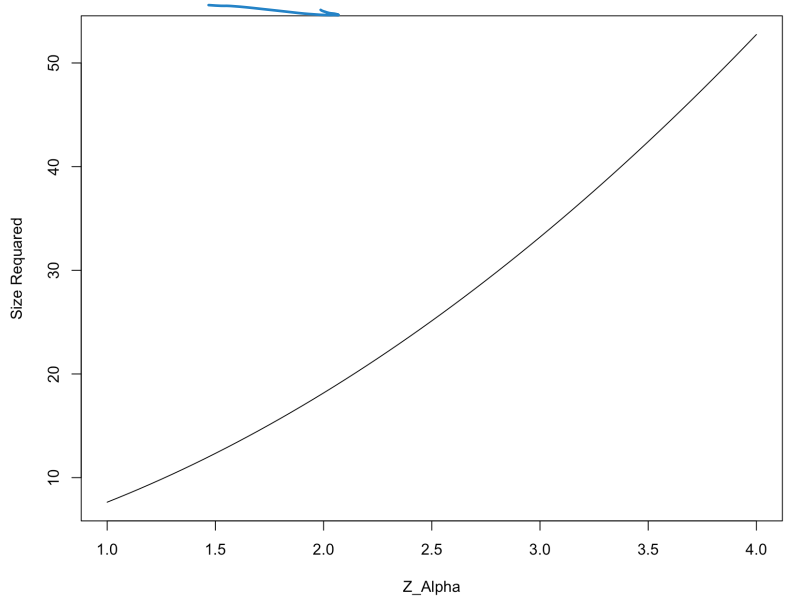
Let the the alternative $\mu = 298$. The sample size n that will suffice according to the formula is,

$$n \geq \frac{\sigma^2(Z_\alpha + Z_{0.8})^2}{\Delta^2} = \frac{9 * (Z_\alpha + 0.84)^2}{4} \tag{27}$$

α (one-side)	β	n
5%	0.2	14
2%	0.2	19
1%	0.2	23
0.1%	0.2	35

seems sensible

stricter



TOP CLASS WORK