

MATH 203: PRINCIPLES OF STATISTICS 1
FALL 2024

SAMPLING DISTRIBUTIONS: EXAMPLES

A *sampling distribution* is simply the probability distribution of a summary statistic random variable: we consider

- a *random sample* of (independent and identically distributed) random variables $X_1, \dots, X_n$ having some pmf $p(x)$ or pdf $f(x)$
- a summary statistic random variable, for example $\bar{X}$, derived from $X_1, \dots, X_n$
- the probability distribution of the summary statistic, calculated from the pmf or pdf of the original variables.

Note: these mathematical calculations are performed **without** considering any observed data (which would be the numerical values $x_1, \dots, x_n$). The calculation is allowing us to specify how the summary statistic should behave if we make assumptions about the distributions of the original random variables.

Example: Suppose random variable X records the number of courses in which a typical student at McGill is registered in a given term, and suppose that the probability distribution of X is specified by the following pmf

x	2	3	4	5
$p(x)$	0.05	0.15	0.6	0.2

That is, the probability that a student is registered in four courses is 0.6 etc. We can compute that the expected number of courses in which a student is registered is

$$E(X) = \sum_{x=1}^4 xp(x) = (2 \times 0.05) + (3 \times 0.15) + (4 \times 0.6) + (5 \times 0.2) = 3.95$$

and for the variance

$$E(X^2) = \sum_{x=1}^4 x^2 p(x) = (2^2 \times 0.05) + (3^2 \times 0.15) + (4^2 \times 0.6) + (5^2 \times 0.2) = 16.15$$

so

$$Var(X) = E(X^2) - \{E(X)\}^2 = 16.15 - 15.6025 = 0.5475$$

so the standard deviation is $\sqrt{0.5475} = 0.7399$.

Suppose we intend to consider $n = 3$ students, and will record the number of courses in which they are registered as random variables X_1, X_2 and X_3 . Before we record the number of courses for each, we can compute the probability distribution of the sample mean random variable

$$\bar{X} = \frac{X_1 + X_2 + X_3}{3}$$

by considering all possible values of the three variables. There are 64 possible configurations, and these are listed in the following table:

Table 1: Possible configurations

Configurations 1 to 32						Configurations 33 to 64					
1	2	3	Sum	Mean	Prob.	1	2	3	Sum	Mean	Prob.
2	2	2	6	2.000000	0.000125	4	2	2	8	2.666667	0.0015
2	2	3	7	2.333333	0.000375	4	2	3	9	3.000000	0.0045
2	2	4	8	2.666667	0.001500	4	2	4	10	3.333333	0.0180
2	2	5	9	3.000000	0.000500	4	2	5	11	3.666667	0.0060
2	3	2	7	2.333333	0.000375	4	3	2	9	3.000000	0.0045
2	3	3	8	2.666667	0.001125	4	3	3	10	3.333333	0.0135
2	3	4	9	3.000000	0.004500	4	3	4	11	3.666667	0.0540
2	3	5	10	3.333333	0.001500	4	3	5	12	4.000000	0.0180
2	4	2	8	2.666667	0.001500	4	4	2	10	3.333333	0.0180
2	4	3	9	3.000000	0.004500	4	4	3	11	3.666667	0.0540
2	4	4	10	3.333333	0.018000	4	4	4	12	4.000000	0.2160
2	4	5	11	3.666667	0.006000	4	4	5	13	4.333333	0.0720
2	5	2	9	3.000000	0.000500	4	5	2	11	3.666667	0.0060
2	5	3	10	3.333333	0.001500	4	5	3	12	4.000000	0.0180
2	5	4	11	3.666667	0.006000	4	5	4	13	4.333333	0.0720
2	5	5	12	4.000000	0.002000	4	5	5	14	4.666667	0.0240
3	2	2	7	2.333333	0.000375	5	2	2	9	3.000000	0.0005
3	2	3	8	2.666667	0.001125	5	2	3	10	3.333333	0.0015
3	2	4	9	3.000000	0.004500	5	2	4	11	3.666667	0.0060
3	2	5	10	3.333333	0.001500	5	2	5	12	4.000000	0.0020
3	3	2	8	2.666667	0.001125	5	3	2	10	3.333333	0.0015
3	3	3	9	3.000000	0.003375	5	3	3	11	3.666667	0.0045
3	3	4	10	3.333333	0.013500	5	3	4	12	4.000000	0.0180
3	3	5	11	3.666667	0.004500	5	3	5	13	4.333333	0.0060
3	4	2	9	3.000000	0.004500	5	4	2	11	3.666667	0.0060
3	4	3	10	3.333333	0.013500	5	4	3	12	4.000000	0.0180
3	4	4	11	3.666667	0.054000	5	4	4	13	4.333333	0.0720
3	4	5	12	4.000000	0.018000	5	4	5	14	4.666667	0.0240
3	5	2	10	3.333333	0.001500	5	5	2	12	4.000000	0.0020
3	5	3	11	3.666667	0.004500	5	5	3	13	4.333333	0.0060
3	5	4	12	4.000000	0.018000	5	5	4	14	4.666667	0.0240
3	5	5	13	4.333333	0.006000	5	5	5	15	5.000000	0.0080

The column ‘Prob’ is the probability of obtaining that configuration, which we compute using independence:

$$P(X_1 = x_1, X_2 = x_2, X_3 = x_3) = P(X_1 = x_1)P(X_2 = x_2)P(X_3 = x_3)$$

for all possible values of x_1, x_2, x_3 . From this table, we deduce that the sample mean random variable can take 10 possible values, namely

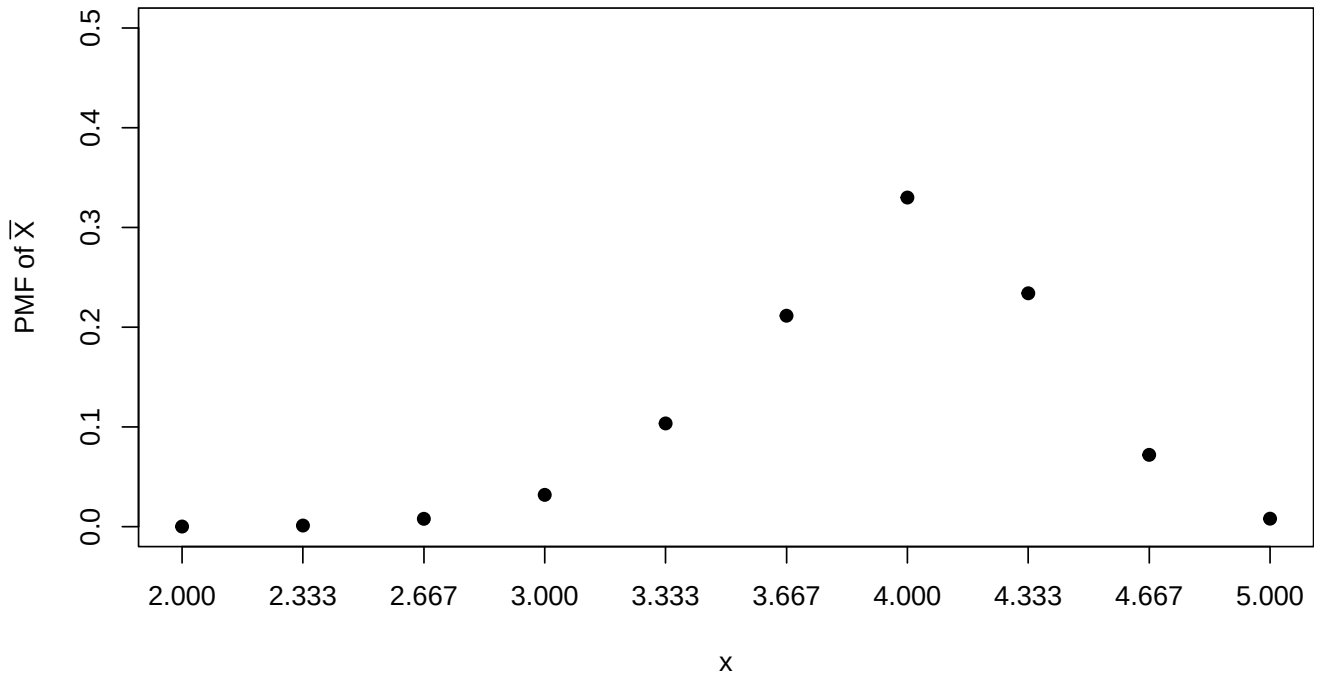
$$\mathcal{X} = \{2.0000, 2.3333, 2.6667, 3.0000, 3.3333, 3.6667, 4.0000, 4.3333, 4.6667, 5.0000\}$$

and we can compute the probability of each possibility simply by adding the probabilities across configurations that give the same mean value.

Table 2: Distribution of Mean random variable

Mean	Probability
2.000000	0.000125
2.333333	0.001125
2.666667	0.007875
3.000000	0.031875
3.333333	0.103500
3.666667	0.211500
4.000000	0.330000
4.333333	0.234000
4.666667	0.072000
5.000000	0.008000

That is, we have that the pmf of $\bar{X}$ is given by the probabilities in this table. We can denote this pmf $p_{\bar{X}}(x)$ (with the subscript indicating which random variable we are dealing with), and plot it in the usual way.



We can treat this as a standard pmf on the set of values $\mathcal{X}$: we can compute the expectation

$$E(\bar{X}) = \sum_{x \in \mathcal{X}} x p_{\bar{X}}(x) = 3.95$$

which is **identical** to the original expectation $E(X)$. For the variance, however, we have by an identical method of calculation that

$$Var(\bar{X}) = 0.1825 = \frac{Var(X)}{3}.$$

What this result tells is that if we assume the original discrete distribution, and we assume that our three students are drawn from this population, then the sample mean that we will get by averaging the number of courses they take will

- take values in $\mathcal{X}$,
- have probability distribution described by pmf $p_{\bar{X}}$
- have expectation equal to the expected number from the original distribution
- have variance equal to the variance from the original distribution divided by three.

We can test this in simulation: suppose we plan to obtain data on 10000 batches of three students. We can use the `sample` function:

```
set.seed(2)
X.sample<-matrix(sample(x,px,size=30000,rep=T),nrow=10000)
Xmean<-apply(X.sample,1,mean)
df.sum$rfreq<-as.numeric(table(Xmean)/10000)
df.sum
```

##		Mean	Probability	rfreq
## 1	2.000000	0.000125	0.0002	
## 2	2.333333	0.001125	0.0013	
## 3	2.666667	0.007875	0.0085	
## 4	3.000000	0.031875	0.0321	
## 5	3.333333	0.103500	0.1056	
## 6	3.666667	0.211500	0.2120	
## 7	4.000000	0.330000	0.3287	
## 8	4.333333	0.234000	0.2305	
## 9	4.666667	0.072000	0.0703	
## 10	5.000000	0.008000	0.0108	

```
mean(Xmean)    #observed sample mean

## [1] 3.947433

var(Xmean)     #observed sample variance

## [1] 0.1870221
```

The relative frequency of obtaining each of the possible 10 values of the mean matches closely the computed pmf, and the sample mean and sample variance are close to their theoretical population versions.

Note: Here we have computed for samples of $n = 3$. We can repeat the same calculation for any n ; however, there will be 4^n possible configurations, and this number grows rapidly, so the exact computation method outlined above is difficult to implement.

Example: Recall the body temperature data from earlier in the course; these are $n = 136$ measurements of body temperature (in Celsius)

```
BodyTemp<-scan('https://www.math.mcgill.ca/dstephens/203/Data/BodyTemp.txt')
n<-length(BodyTemp)
c(mean(BodyTemp), var(BodyTemp), sd(BodyTemp))

## [1] 36.8076923 0.1659869 0.4074149
```

For these observed data, we have the sample summaries

$$\bar{x} = 36.8076923 \quad s^2 = 0.1659869 \quad s = 0.4074149$$

Suppose we plan to carry out a similar measurement experiment by collecting data on $n = 136$ patients in each of ten hospitals, where we believe that the body temperature of all patients will follow a $Normal(\mu, \sigma^2)$ distribution with $\mu = 37$ (Celsius) and $\sigma = 0.40$. We might observe the following sample means once we have collected the data:

```
## [1] 36.97070 37.05469 36.99317 37.07448 36.97414 36.98340 36.94689 37.06964
## [9] 36.97027 37.02533
```

As we see, there is some variation in the observed sample means across hospitals, which merely reflects sampling variability. We can prove mathematically that the sample mean random variable has a distribution which is also Normal, that is

$$\bar{X} \sim Normal\left(\mu, \frac{\sigma^2}{n}\right)$$

and we have thus characterized the probability distribution of $\bar{X}$. Note that if we consider the Z -transform method, we have that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \frac{(\bar{X} - \mu)}{\sigma} \sim Normal(0, 1)$$