National Institute of Environmental Health Sciences

Summer 2023 Biostatistics and Bioinformatics Short Courses

Statistical Inference

Module 5: Variability
Module 6: Distribution
Module 7: Asymptotic

Komlan Atitey, PhD

Biostatistics and Computational Biology Branch

Outline

- Introduction – Example of dataset
- Variability
  - Population variance
  - Sample variance
  - Standard error
- Distribution
  - Binomial
  - Normal
  - Poisson
- Asymptotic
  - Law of large numbers
  - Central limit theory
  - Confidence interval
- Quiz

Introduction – RNA-seq Data

- **Example of dataset: RNA-seq**
  - RNA-seq (RNA-sequencing) is a technique that can examine the quantity and sequences of RNA in a sample using next-generation sequencing (NGS)
  - It analyzes the transcriptome, indicating which of the genes encoded in our DNA are turned on or off and to what extent.

- **Simulate RNA-seq Data**
  - We may simulate RNA data of the airway data from Himes et al (2014)
  - The airway package contains an example of dataset from an RNA-Seq experiment of read counts per gene for airway smooth muscles
    - ✓ Each row of the airway count data is the gene (64,102 genes)
    - ✓ Each column of the airway count data is a sample (8 samples)

- **Simulation in R**

```
> library(airway)
> data("airway")
> # what the RNAseq data looks like?
> data.airway <- assays(airway)$counts
> # dimension of the RNAseq data
> dim(data.frame(data.airway)) # There are 58395 genes in our samples
[1] 64102     8
> # read the first 5 raws of the RNAseq data
> head(data.airway[c(1:5),])
                 SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
ENSG00000000003        679        448        873        408       1138       1047        770        572
ENSG00000000005          0          0          0          0          0          0          0          0
ENSG00000000419        467        515        621        365        587        799        417        508
ENSG00000000457        260        211        263        164        245        331        233        229
ENSG00000000460         60         55         40         35         78         63         76         60
```

Variability

- Variability refers to how spread out a set of data is.
- Variability gives you a way to describe how much data sets vary and allows you to use statistics to compare your data to other sets of data.
- Example of 4 main ways to describe variability in a data set are:
  - **Range**

    The range is the amount between your smallest and largest item in the set.
  - **Interquartile range**

    The interquartile range is almost the same as the range, only instead of stating the range for the whole data set, you're giving the amount for the "middle fifty".
  - **Variance**

    The variance of a data set gives you a rough idea of how spread out your data is.
  - **Standard deviation**

    The standard deviation tells you how tightly your data is clustered around the mean

Population Variance

- The mean μ of a data set is the population mean which represents the sum of all the data (X) divided by the count.

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

- The variance (σ^2) of a random variable is a measure of spread

- The variance of a random variable X, with mean E[X] = μ is defined as

$$\sigma^2 = Var(X) = E[(X-\mu)^2] = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

- The standard deviation is defined as the square root of the variance.

$$\sigma = SD(X) = \sqrt{Var(X)}$$

```
> ############ compute population variance and standard deviation
> x1 <- 1:100
> population.variance <- sum((x1 - mean(x1))^2)/100
> population.variance
[1] 833.25
> population.sd <- sqrt(population.variance)
> population.sd
[1] 28.86607
```

Sample Variance

- A sample is a set of observations that are pulled from a population and can completely represent it.
- Sample variance is used to calculate the variability in a given sample.
- The sample variance (S^2) is the average square of distance of the observed data minus the sample mean

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n\text{-}1}$$

  - The denominator ($n-1$) will make unbiased estimator of population variance
  - It is also a random variable, and has a population distribution with expected value equals the population variance

$$E(S^2) = Var(X)$$

- The sample standard deviation (S) is the square root of the sample variance

```
> ############ compute sample variance
> x1 <- 1:100
> sample.variance <- sum((x1 - mean(x1))^2)/99
> sample.variance
[1] 841.6667
> sample.variance <- var(x1)
> sample.variance
[1] 841.6667
```

Standard Error of the Mean

- The standard error of the mean, or simply standard error, indicates how different the population mean is likely to be from a sample mean.

$$E(\bar{X}) = \mu$$

- It tells you how much the sample mean would vary if you were to repeat a study using new samples from within a single population.

- The variance of sample mean

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

❑ Why standard error matters

- In statistics, data from samples is used to understand larger populations. Standard error matters because it helps you estimate how well your sample data represents the whole population.

❑ Standard error vs standard deviation

Standard error and standard deviation are both measures of variability:

- ✓ The standard deviation describes variability within a single sample.
- ✓ The standard error estimates the variability across multiple samples of a population.

❑ Standard error formula

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}} \qquad \begin{cases} SE & \text{is standard error} \\ \sigma & \text{is population standard deviation} \\ n & \text{is the number of elements in the sample} \end{cases}$$

Example

Calculate the variance, standard deviation and standard error of the gene expression for one gene

- **Calculate sample variance**

```
> ############ Calculate sample variance
> first.row <- data.airway[1,] # first row data
> first.row
SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
       679        448        873        408       1138       1047        770        572
> n <- ncol(data.airway) # sample size
> s.mu <- mean(first.row) # sample mean
> s.var <- sum((first.row - s.mu)^2)/(n-1) #variance
> s.var
[1] 71235.27
> s.var <- var(first.row) # use function var to calculate sample variance
> s.var
[1] 71235.27
```

- **Calculate sample standard deviation**

```
> ############Calculate sample standard deviation
> s.sd <- sd(first.row) ### standard deviation (1)
> s.sd
[1] 266.8994
> s.sd <- sqrt(var(first.row)) ### square root of variance (2)
> s.sd
[1] 266.8994
> sd(first.row) == sqrt(var(first.row )) ### compare (1) and (2)
[1] TRUE
```

- **Calculate sample standard error**

```
> ############ Calculate sample standard error
> n <- ncol(data.airway) ### sample size
> s.sde <- sd(data.airway[1,])/sqrt(n) ### standard error
> s.sde
[1] 94.36317
```

R function for performing row-wise or column-wise calculations

We can use one function to calculated the mean, variance and standard deviation for all variables in a data matrix in one function instead of using for loop

- **Number of rows**

```
> ############ number of rows
> nrow(data.airway) ### number of genes (rows)
[1] 64102
```

- **Mean**

```
> ############ mean
> mu <- rowMeans(data.airway) ### mean
> head(mu)
ENSG00000000003 ENSG00000000005 ENSG00000000419 ENSG00000000457 ENSG00000000460 ENSG00000000938
        741.875           0.000         534.875         242.000          58.375           0.375
```

- **Variance**

```
> ############ variance
> var <- rowVars(data.airway) ### variance
> head(var)
[1] 7.123527e+04 0.000000e+00 1.833898e+04 2.284286e+03 2.311250e+02 5.535714e-01
```

- **Standard deviation**

```
> ############ standard deviation
> sd <- rowSds(data.airway) ### standard deviation
> head(sd)
[1] 266.8993590   0.0000000 135.4214981  47.7942017  15.2027958   0.7440238
```

Distributions

In statistics, a distribution is a function that shows the possible values for a variable and how often they occur within a given dataset. It enables you to calculate the probability of certain outcomes occurring, and to understand how much variation there is within your dataset.

- Discrete vs Continuous

  A discrete random variable has a finite number of possible values. A continuous random variable could have any value (usually within a certain range)

- Binomial
  - Gender,
  - Disease status, etc.

- Normal
  - Body Mass Index (BMI), defined as the ratio of individual mass (in kilograms) to the square of the associated height (in meters), is one of the most widely discussed and utilized risk factors in medicine and public health, given the increasing obesity worldwide and its relation to metabolic disease.
  - Statistically, BMI is a composite random variable, since human weight (converted to mass) and height are themselves random variables.

- Poisson
  - Read counts from High-throughput sequencing, etc.

Binomial Distribution *Bin(n,p)*

- Let's consider performing *n* independent trials, where each trial can result in a "success" with probability p and a "failure" with probability $1-p$

- If *X* represents the number of successes that occur in the *n* trials, then *X* is said to be a binomial random variable with parameters (*n*, *p*).

  - The probability mass function (PMF) of a binomial random variable *X* with parameters *n* and *p* is given by

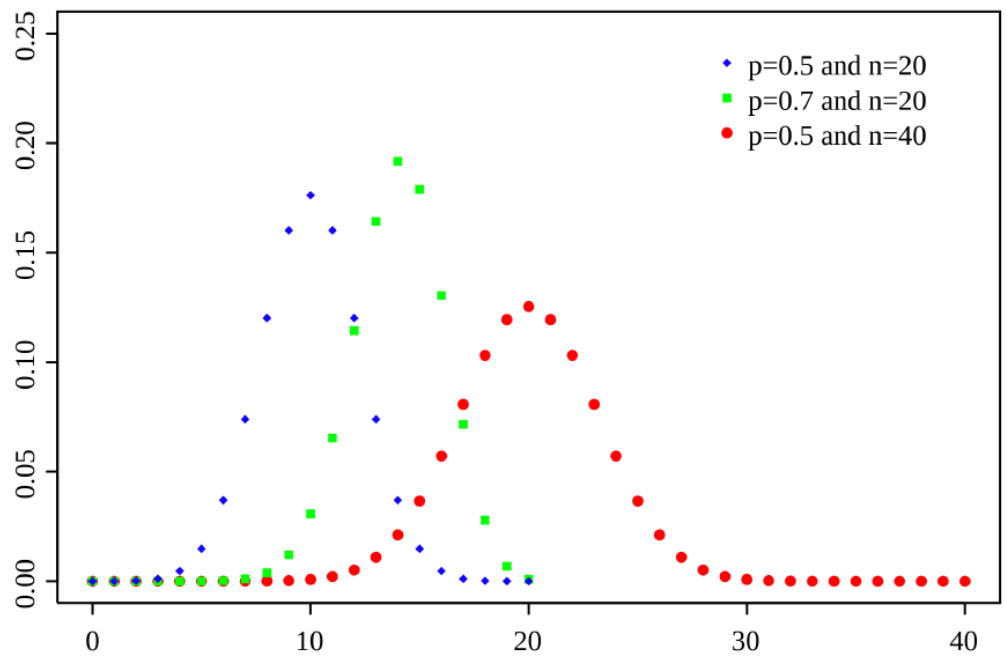
$$f(X=x,n,p)=f(x)=\binom{n}{x}p^x(1-p)^{n-x} \qquad \text{For } x=0,1,2,\cdots,n$$

where $\binom{n}{x}=\frac{n!}{x!\,(n-x)!}$ is the number of combination which represent the binomial coefficient

  - The mean: $E(X)=np$

  - The variance: $\mathrm{Var}(X)=np(1-p)$

- Bernoulli distribution is a special case of binomial when n = 1



Normal Distribution $N(\mu, \sigma^2)$

- The Normal Distribution or Gaussian Distribution is defined by the probability density function for a continuous random variable in a system.

  - The probability density function of normal distribution is given by

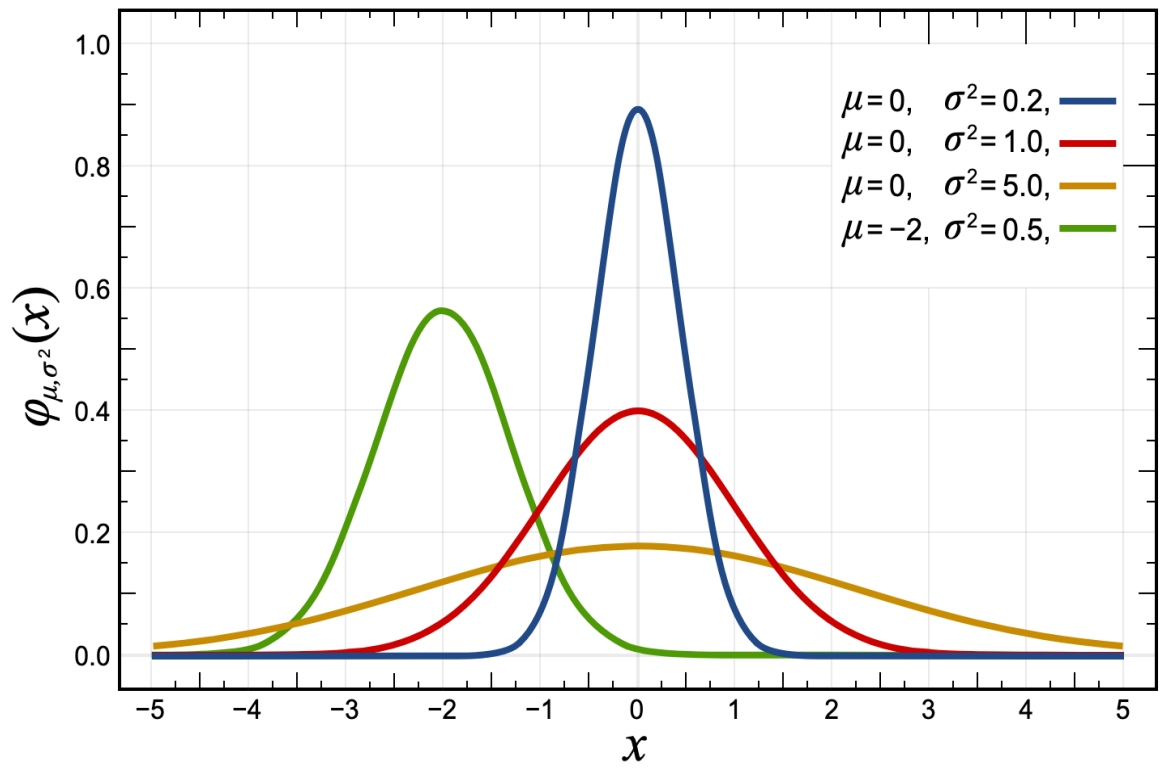
$$f(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Mean: $E(X) = \mu$
Variance: $Var(X) = \sigma^2$ } Notation $X \sim N(\mu, \sigma^2)$

  - Special case: μ=0, σ^2=1

    The standard normal distribution is a normal distribution with a mean of zero and standard deviation of 1

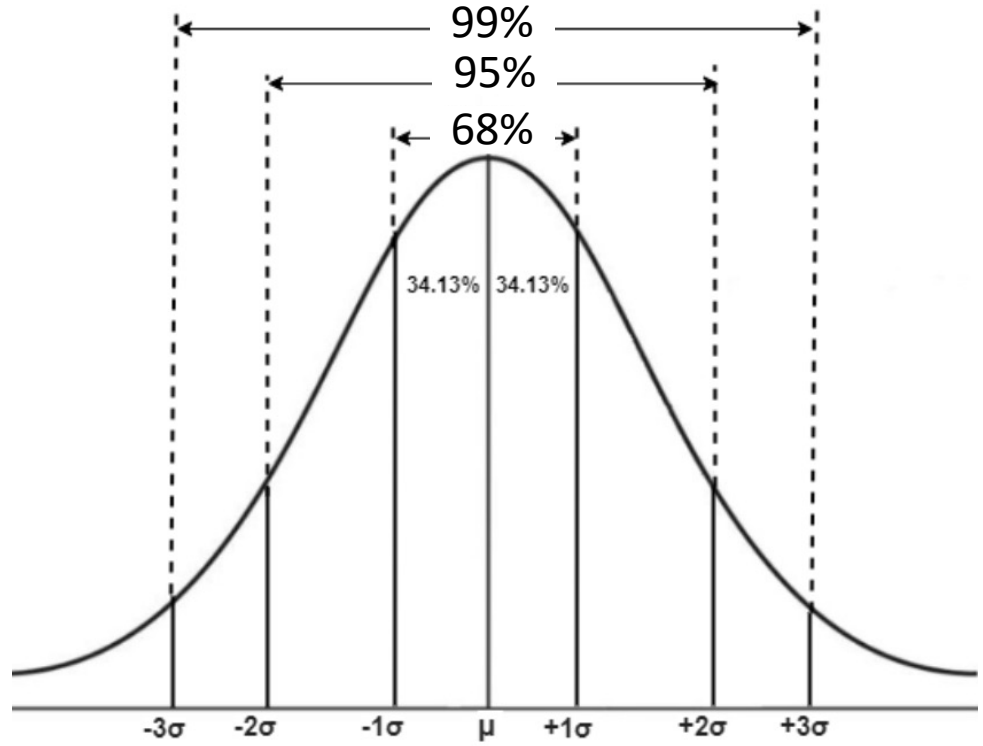
$$Z \sim N(0,1)$$



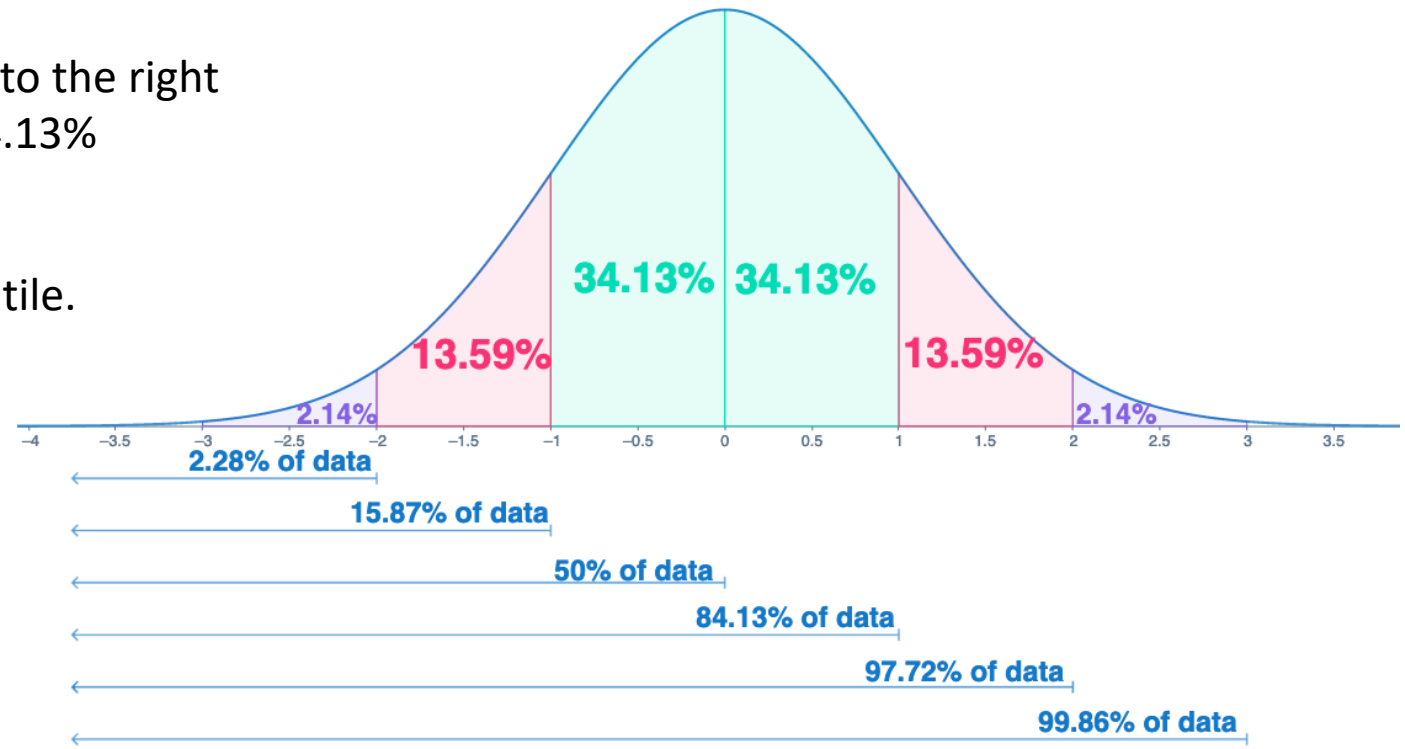
Normal Quantiles

- A quantile determines how many values in a distribution are above or below a certain limit.

- μ and σ determine respectively the center and spread of the distribution
  - 68% of the area under the curve lies between $(\mu - \sigma, \mu + \sigma)$
  - 95% of the area under the curve lies between $(\mu - 2\sigma, \mu + 2\sigma)$
  - 99% of the area under the curve lies between $(\mu - 3\sigma, \mu + 3\sigma)$



- A percentile is a measure used in statistics indicating the value below which a given percentage of observations in a group of observations fall.
  - For 1 standard deviation above the mean, that is to the right of the mean, find the percentile by adding the 34.13% above the mean to the 50% to get 84.13%.
  - So, 1 standard deviation is about the 84th percentile.



- Z score

  A Z-Score is a statistical measurement of a score's relationship to the mean in a group of scores

$$Z = \frac{x - \mu}{\sigma}$$

Example

- z-score provides how many standard deviations away a value is from the mean.

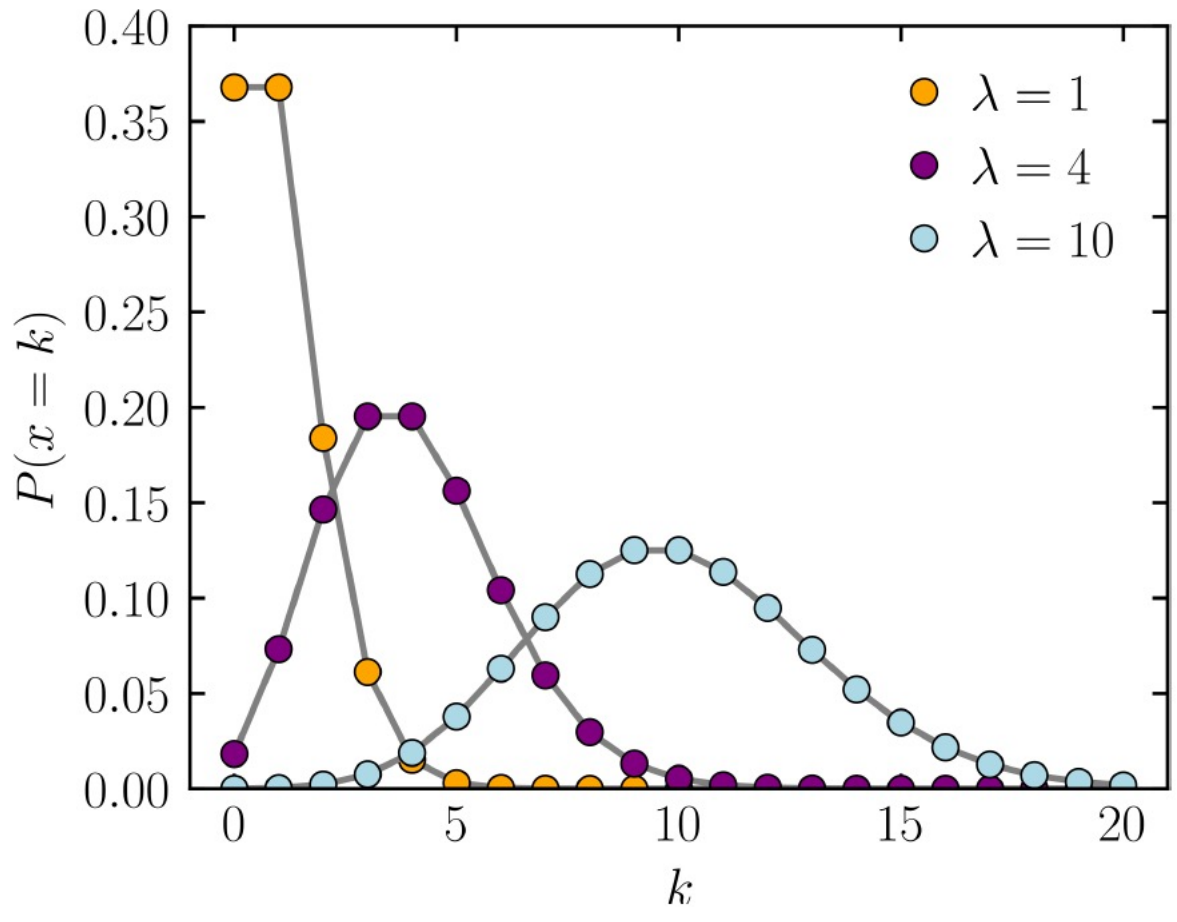
```
> ############################################## z score
> data <- c(8, 7, 7, 10, 13, 14, 15, 16, 18) # generate a data
> z_scores <- (data-mean(data))/sd(data) # compute z score of each value
> z_scores
[1] -0.9701425 -1.2126781 -1.2126781 -0.4850713  0.2425356  0.4850713  0.7276069  0.9701425  1.4552138
> plot(z_scores, type="o", col="red") # plot
```

Z score percentile Normal Distribution Table

Percentile	z-Score	Percentile	z Score	Percentiles	z - Score
1	-2.326	34	-0.412	67	0.44
2	-2.054	35	-0.385	68	0.468
3	-1.881	36	-0.358	69	0.496
4	-1.751	37	-0.332	70	0.524
5	-1.645	38	-0.305	71	0.553
6	-1.555	39	-0.279	72	0.583
7	-1.476	40	-0.253	73	0.613
8	-1.405	41	-0.228	74	0.643
9	-1.341	42	-0.202	75	0.674
10	-1.282	43	-0.176	76	0.706
11	-1.227	44	-0.151	77	0.739
12	-1.175	45	-0.126	78	0.772
13	-1.126	46	-0.1	79	0.806
14	-1.08	47	-0.075	80	0.842
15	-1.036	48	-0.05	81	0.878
16	-0.994	49	-0.025	82	0.915
17	-0.954	50	0	83	0.954
18	-0.915	51	0.025	84	0.994
19	-0.878	52	0.05	85	1.036
20	-0.842	53	0.075	86	1.08
21	-0.806	54	0.1	87	1.126
22	-0.772	55	0.126	88	1.175
23	-0.739	56	0.151	89	1.227
24	-0.706	57	0.176	90	1.282
25	-0.674	58	0.202	91	1.341
26	-0.643	59	0.228	92	1.405
27	-0.613	60	0.253	93	1.476
28	-0.583	61	0.279	94	1.555
29	-0.553	62	0.305	95	1.645
30	-0.524	63	0.332	96	1.751
31	-0.496	64	0.358	97	1.881
32	-0.468	65	0.385	98	2.054
33	-0.44	66	0.412	99	2.326

Poisson Distribution Pois(λ)

- A Poisson distribution is a discrete probability distribution, meaning that it gives the probability of a discrete (i.e., countable) outcome.
- For Poisson distributions, the discrete outcome is the number of times an event occurs.
- The Poisson distribution has only one parameter, λ (lambda), which is the mean number of events.
  - ***PMF*** $f(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
  - ***Mean*** $E(X) = \lambda$
  - ***Variance*** $Var(X) = \lambda$



- You can use a Poisson distribution to predict or explain the number of events occurring within a given interval of time or space.
- The interval can be any specific amount of time or space, such as 10 days or 5 square inches.
- Poisson approximates the binomial distribution when *n* is large and *p* is small

Examples

Problem

If there are twelve cars crossing a bridge per minute on average, find the probability of having seventeen or more cars crossing the bridge in a particular minute.

Solution

The probability of having *sixteen or less* cars crossing the bridge in a particular minute is given by the function ppois.

```
> ppois(16, lambda=12)   # lower tail
[1] 0.898709
```

Hence the probability of having seventeen or more cars crossing the bridge in a minute is in the *upper tail* of the probability density function.

```
> ppois(16, lambda=12, lower=FALSE)   # upper tail
[1] 0.101291
```

Examples

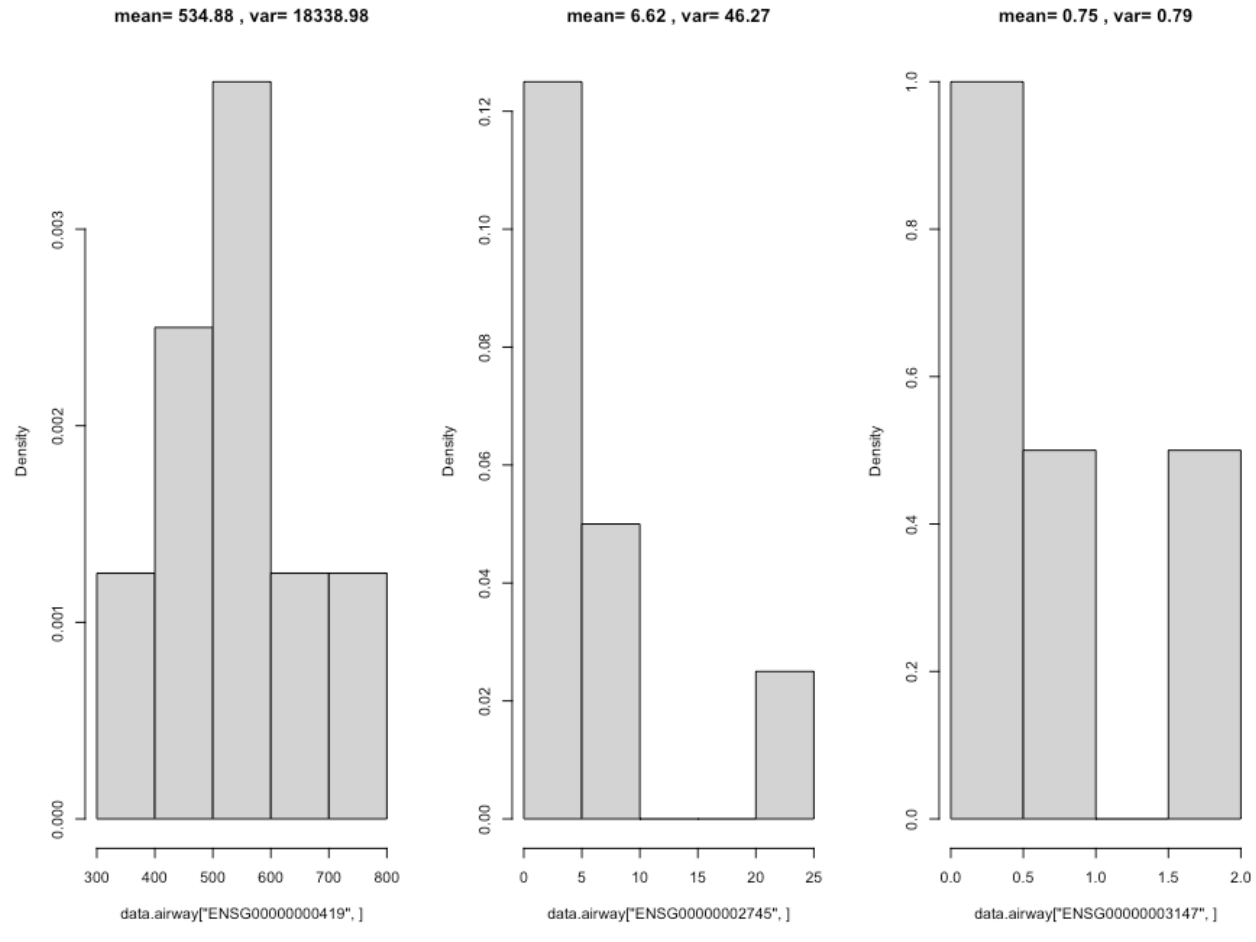
Consider 3 genes' expression value from RNA-seq data

```
> ############################################## gene expression across samples
> data.airway["ENSG00000000419",]
SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
       467        515        621        365        587        799        417        508
> data.airway["ENSG00000002745",]
SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
         4          6         22         10          2          1          5          3
> data.airway["ENSG00000003147",]
SRR1039508 SRR1039509 SRR1039512 SRR1039513 SRR1039516 SRR1039517 SRR1039520 SRR1039521
         0          0          0          1          1          2          2          0
```

Histogram of Gene Expression

- Histogram of gene expression for 3 genes

```
par(mfrow=c(1,3))
hist(data.airway["ENSG00000000419",], freq = FALSE,
     main = paste("mean=", round(mean(data.airway["ENSG00000000419",]),2),
                  ", var=", round(var(data.airway["ENSG00000000419",]),2)))

hist(data.airway["ENSG00000002745",], freq = FALSE,
     main = paste("mean=", round(mean(data.airway["ENSG00000002745",]),2),
                  ", var=", round(var(data.airway["ENSG00000002745",]),2)))

hist(data.airway["ENSG00000003147",], freq = FALSE,
     main = paste("mean=", round(mean(data.airway["ENSG00000003147",]),2),
                  ", var=", round(var(data.airway["ENSG00000003147",]),2)))
```



Negative Binomial NB(μ, θ)

- A negative binomial distribution (also called the Pascal distribution) is a discrete probability distribution for random variables in a negative binomial experiment.

- The negative binomial distribution combines the sampling variability of a Poisson and biological variability, is a more appropriate distribution to model biological experiment

- ***PMF*** $f(X = x) = \frac{\Gamma(\theta + x)}{x!\,\Gamma(\theta)} \left(\frac{\theta}{\theta + \mu}\right)^{r} \left(\frac{\mu}{\theta + \mu}\right)^{x}$ For x failures given θ successes

- ***Mean*** $E(X) = \mu$

- ***Variance*** $Var(X) = \mu + \frac{\mu^2}{\theta}$

- When θ goes to infinity, Poisson distribution converges to negative binomial

Asymptotic

- In statistics, asymptotic theory, or large sample theory, is a framework for assessing properties of estimators and statistical tests.
- Asymptotic refers to the behavior of estimators as the sample size goes to infinity
- Why asymptotic statistics?
  - It enable us to find approximate tests and confidence regions.
  - approximations can be used theoretically to study the quality (efficiency) of statistical procedures

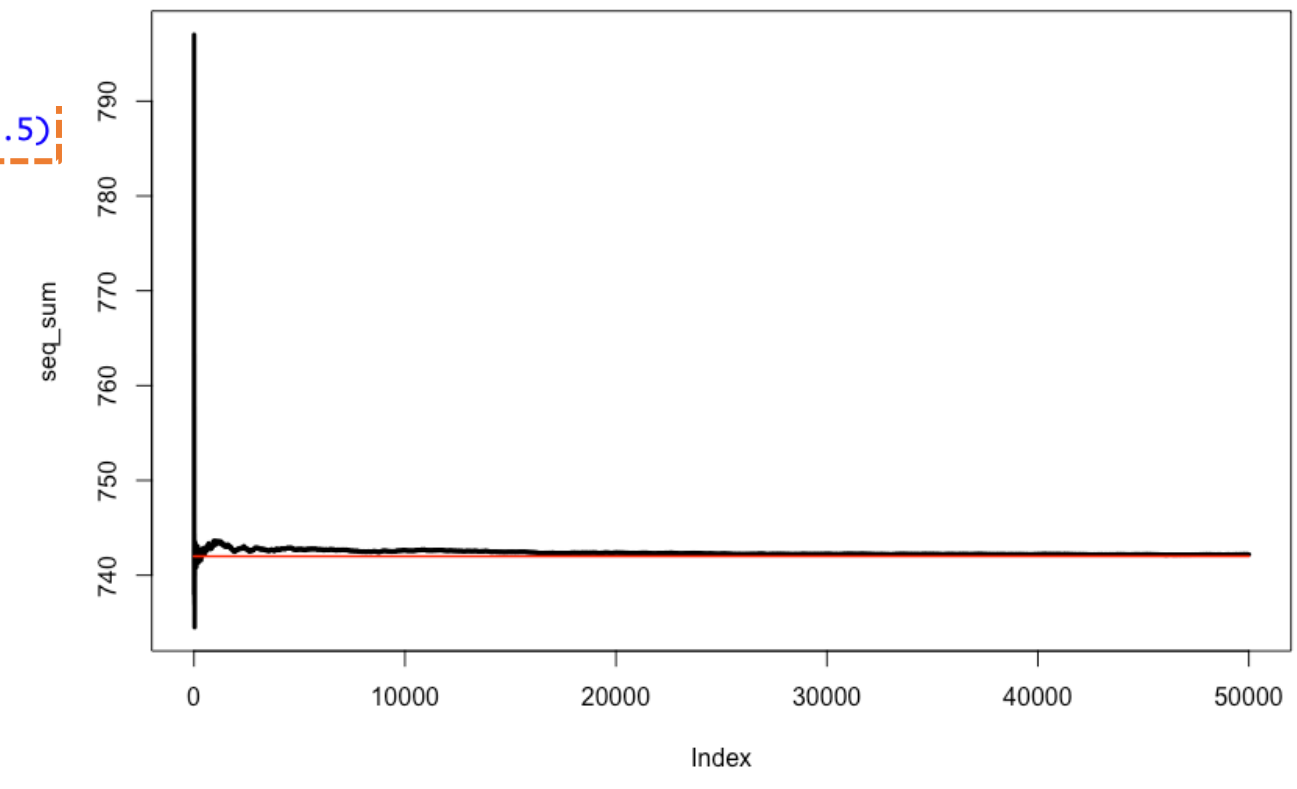
Law of Large Number (LLN)

- It states that if you repeat an experiment independently a large number of times and average the result, what you obtain should be close to the expected value.
- The sample mean $(\bar{X})$ will converge to the population mean (μ) if the number of trials or sample size (n) is large enough.
- To empirically estimate the expected value of a random variable, one repeatedly measures observations of the variable and computes the arithmetic mean of the results.

Example

```
> ### assume a gene expression follow a Poisson distribution with population mean 742
> N <- 50000
> lambda <- round(mean(data.airway[1,]))
> simu <- rpois(N, lambda)
> simu_1 <- sample(simu, 10)
> mean(simu_1)
[1] 753.2
> simu_2 <- sample(simu, 1000)
> mean(simu_2)
[1] 741.239
> simu_3 <- sample(simu, 10000)
> mean(simu_3)
[1] 742.1075
> ### plot the results
> seq_sum <- NULL
> for (i in 1:N) {
+   seq_sum[i] <- sum(simu[1:i])/i
+ }
> plot(seq_sum, type = "l",
+      xlabel = "Observations",
+      ylabel = "", lwd =3)
> lines(c(0,N), c(lambda, lambda), col="red", lwd = 1.5)
```

rpois() function in R Language is used to compute random density for poisson distribution.



The Central Limit Theorem (CLT)

- The Central Limit Theorem states that the sampling distribution of the sample means approaches a normal distribution as the sample size gets larger — *no matter what the shape of the population distribution*.

- Let $X_1, X_2, \cdots, X_n$ denote a random sample of n independent observations from a population with overall expected value μ and finite variance, and denote the sample mean of that sample – itself a random variable by $\bar{X}_n$.

- Then, the distribution of sample mean $\bar{X}_n$ is approximately a normal distribution $N(\mu, \frac{\sigma^2}{n})$ if the sample size (n) is large enough.

  - ❑ The mean of the sampling distribution will be equal to the mean of the population distribution

  - ❑ The standard deviation of the sampling distribution will be equal to the standard deviation of the population distribution divided by the sample size

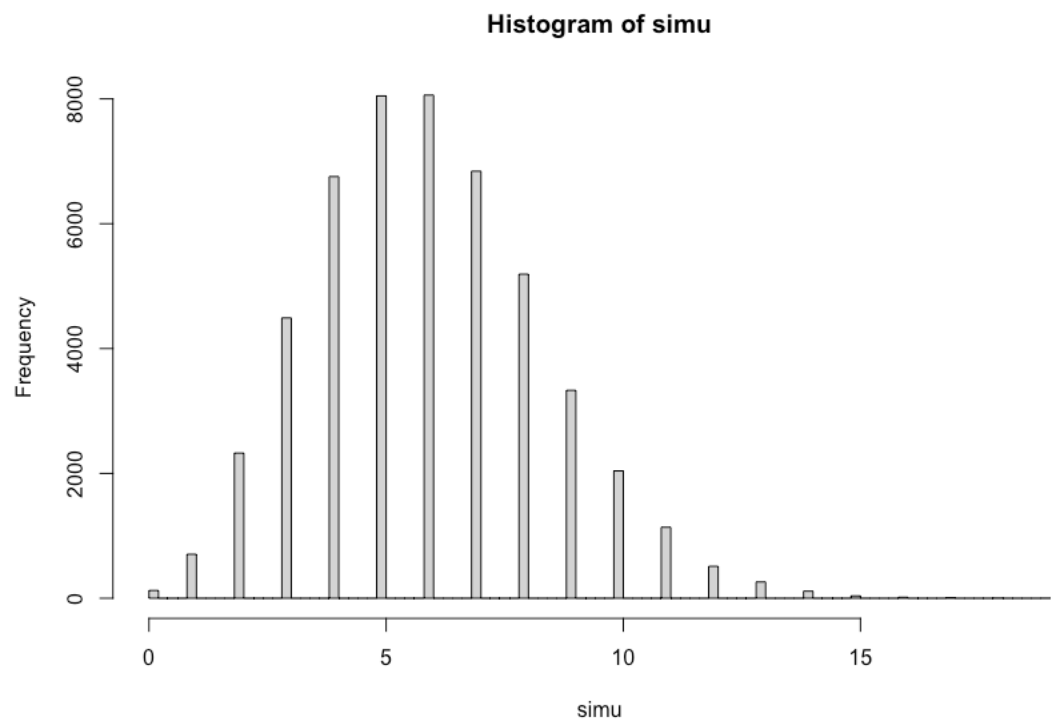
- The theorem is a key concept in probability theory because it implies that probabilistic and statistical methods that work for normal distributions can be applicable to many problems involving other types of distributions.

Example

- Assume the true distribution of a gene expression follows a Poisson distribution with population mean equals to 6

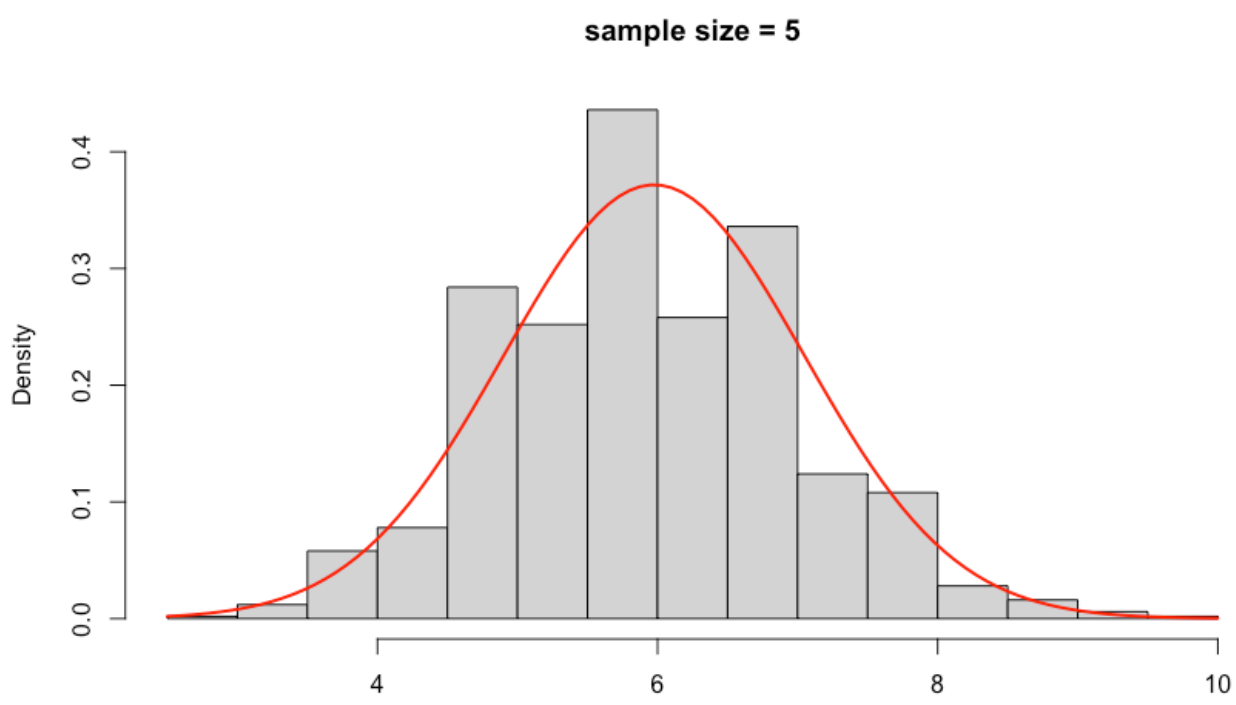
- Assume population size is 50000

```
############################################## distribution of gene expression
set.seed(123) # make this example reproducible

# assumea gene expresion is Poisson distributed with population mean equals to 6
# (close to the mean of gene 16 "ENSG00000001626")
N <- 50000
lambda <- round(mean(data.airway[16,]))
simu <- rpois(N,lambda)

############
############ create histogram to visualize distribution of gene expression
hist(simu, 100)
```



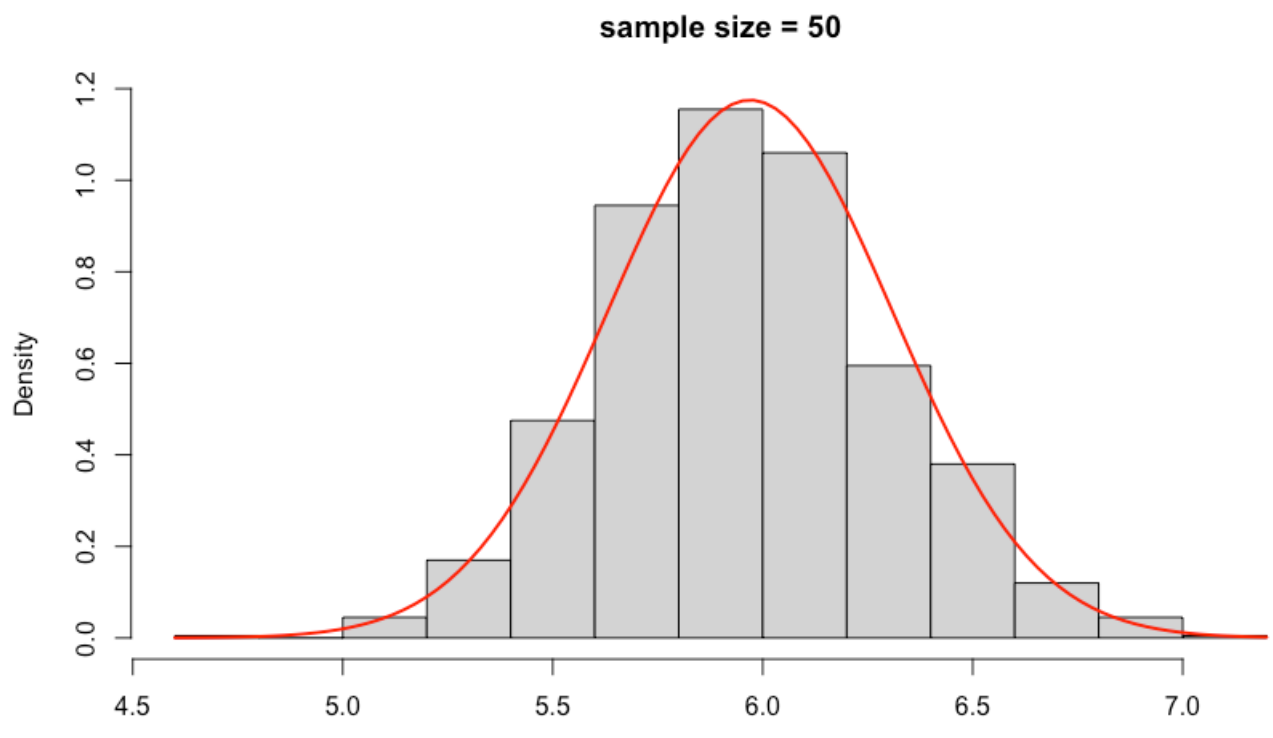
When sample size n=5

```
> ############################################## 1000 random samples of size 5
> n <- 1000
> sample_1 <- NULL
> for (i in 1:n) {
+   sample_1[i] <- mean(sample(simu, 5, replace = TRUE))
+ }
> mean(sample_1)
[1] 5.9568
> sd(sample_1)
[1] 1.096164
> ############
> ############ create histogram to visualize sampling distribution of sample means
> hist(sample_1, xlab="", main = "sample size = 5", freq = FALSE)
> curve(dnorm(x, mean=mean(sample_1), sd = sd(sample_1)), col="red", lwd=2, add=TRUE, yaxt="n")
```



When sample size n=50

```
> ############################################## 1000 random samples of size 50
> n <- 1000
> sample_1 <- NULL
> for (i in 1:n) {
+   sample_1[i] <- mean(sample(simu, 50, replace = TRUE))
+ }
> mean(sample_1)
[1] 5.96984
> sd(sample_1)
[1] 0.3394844
> ############
> ############ create histogram to visualize sampling distribution of sample means
> hist(sample_1, xlab="", main = "sample size = 50", freq = FALSE)
> curve(dnorm(x, mean=mean(sample_1), sd = sd(sample_1)), col="red", lwd=2, add=TRUE, yaxt="n")
```

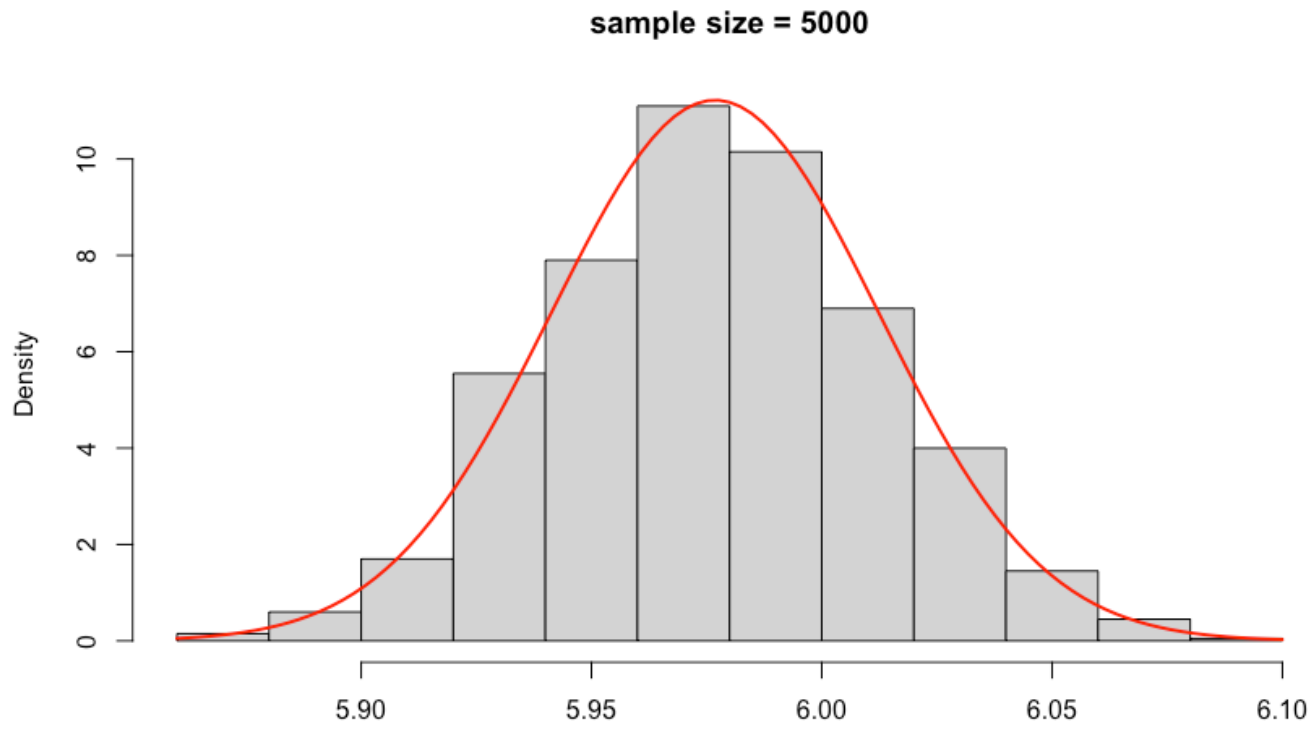


When sample size n=5000

```
> ############################################## 1000 random samples of size 5000
> n <- 1000
> sample_1 <- NULL
> for (i in 1:n) {
+   sample_1[i] <- mean(sample(simu, 5000, replace = TRUE))
+ }
> mean(sample_1)
[1] 5.976832
> sd(sample_1)
[1] 0.03555095
> ############
> ############ create histogram to visualize sampling distribution of sample means
> hist(sample_1, xlab="", main = "sample size = 5000", freq = FALSE)
> curve(dnorm(x, mean=mean(sample_1), sd = sd(sample_1)), col="red", lwd=2, add=TRUE, yaxt="n")
```

Sample size	Sample mean
5	5.9568
50	5.96984
5000	5.976832

Population mean = 6



Confidence Intervals

- A confidence interval, in statistics, refers to the probability that a population parameter will fall between a set of values for a certain proportion of times. Analysts often use confidence intervals that contain either 95% or 99% of expected observations.

- The confidence interval is obtained by adding and subtracting the margin of error from the sample mean. This result is the upper limit and the lower limit of the confidence interval.

❑ ***Confidence interval (Ci)***

$$Ci = \text{Sample mean} \pm \text{Margin of error}$$

❑ ***Margin of error (Me)***

$$Me = \frac{Z\ value \times Population\ standard\ deviation}{\sqrt{Sample\ size}}$$

❑ ***General*** $(\alpha\%)$ ***Ci***

$$Ci = \text{Sample mean} \pm z_{(1-\alpha/2)} \cdot \frac{\sigma}{\sqrt{n}}$$

$z_{(1-\alpha/2)}$ is the $1 - \alpha/2$ standard normal quantile

α is the significance level used to compute the confidence interval

Example: Compute ci for $\alpha = 0.05$

- Normal approximation

$$Ci = \bar{X} \pm z_{0.975}\frac{\sigma}{\sqrt{n}}$$

se

For a normal distributed random variable, the function **qnorm()** in R, compute the probabilities from known bounding values.

```
> ############################################## confidence interval
> mu <- mean(data.airway["ENSG00000002079",])
> mu
[1] 1.5
> var <- var(data.airway["ENSG00000002079",])
> var
[1] 2.571429
> se <- sd(data.airway["ENSG00000002079",])/sqrt(n)
> se
[1] 0.05070926
> ############
> ########### normal approximation
> mu + c(-1,1)*qnorm(0.975)*se
[1] 1.400612 1.599388
```

- Gamma approximation

$$Ci = \bar{X} \pm \gamma_{0.975}\frac{\sigma}{\sqrt{n}}$$

$\gamma_{0.975}$ is the $(1 - {}^{\alpha}/_{2})^{th}$ gamma quantile

The **qgamma()** function creates a quantile plot of a gamma distribution

```
> ########### gamma approximation
> theta <- mu^2/(var-mu) # dispersion
> a <- theta*n # shape parameter
> b <- theta*n/mu # # rate parameter
> mu + c(-1,1)*qgamma(0.975, shape = a, rate = b)*se
[1] 1.420649 1.579351
```

Sample mean vs Population mean

- The sample mean is the arithmetic average computed using samples or random values taken from the population. It is evaluated as the sum of all the sample variables divided by the total number of variables.

- A population mean is the average computed from the entire group, distribution, or population. It is derived by dividing the aggregate of all the population variables by the total number of variables in the population.

❑ ***Meaning***

- ✓ Sample mean: Only considers a selected number of observations—drawn from the population data.
- ✓ Population mean: Considers all the observations in the population—to compute the average value.

❑ ***Number of Observations***

- ✓ Sample mean: Sample size is small, and is represented by '**n**'
- ✓ Population mean: Population size is large, and is denoted by '**N**'

❑ ***Calculation***

- ✓ Sample mean: $\bar{x} = \frac{\sum_i x_i}{n}$
- ✓ Population mean: $\mu = \frac{\sum X}{N}$

❑ ***Standard Deviation***

- ✓ Sample mean: The standard deviation derived from a sample mean is represented by '**s**'
- ✓ Population mean: The standard deviation evaluated from the population mean is represented by '$\boldsymbol{\sigma}$'.

By law of large numbers (LLN), sample mean will converge to population mean when sample size is large enough.

Quiz 3

- Brain volume for adult women is normally distributed with a mean of about 1,100 cc (cube centimeter) for women with a standard deviation of 75 cc. What brain volume represents the 95th percentile?

  - approximately 1247
  - approximately 1223
  - approximately 1175
  - approximately 977

Answer 3

- Calculate using formula
  - This is the population distribution
  - Recall 1.645 is the 95th percentiles of the standard normal distribution
  - $\mu + Z * \sigma$ = 1100+1.645*75 = 1223

- use R function

```
> ########################################## Quiz 3
> qnorm(0.95, mean = 1100, sd = 75)
[1] 1223.364
```

Quiz 4

- Refer to the previous question. Brain volume for adult women is normally distributed with a mean of about 1,100 cc for women with a standard deviation of 75 cc. Consider the sample mean of 100 random adult women from this population. What is the 95th percentile of the distribution of the sample mean?

  - Approximately 1112
  - Approximately 1088
  - Approximately 1110
  - Approximately 1115

Answer 4

- Here, we have samples randomly chosen from the population, with sample size n=100

- By CLT, $\mathrm{E}(\bar{X}) = \mu = 1100$

- Standard deviation of sample mean is $\mathrm{sd}(\bar{X}) = \frac{\sigma}{\sqrt{n}} = \frac{75}{10}$

- Recall 1.645 is the 95th percentiles of the standard normal distribution

- $\mu + Z * \sigma$ = 1100+1.645∗(75/10) = 1112.338

Quiz 1

- What is the variance of the distribution of the average an IID draw of n observations from a population with mean μ and variance σ2?
  - $\frac{\sigma^2}{n}$
  - σ^2
  - $2\sigma n\sqrt{}$
  - σn

Answer 1

$$Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right)$$

$$Var(\bar{X}) = \frac{1}{n^2} Var(X_1) + \cdots + \frac{1}{n^2} Var(X_n)$$

$$Var(\bar{X}) = \frac{1}{n^2} n\sigma^2$$

$$Var(\bar{X}) = \frac{\sigma^2}{n}$$

Quiz 2

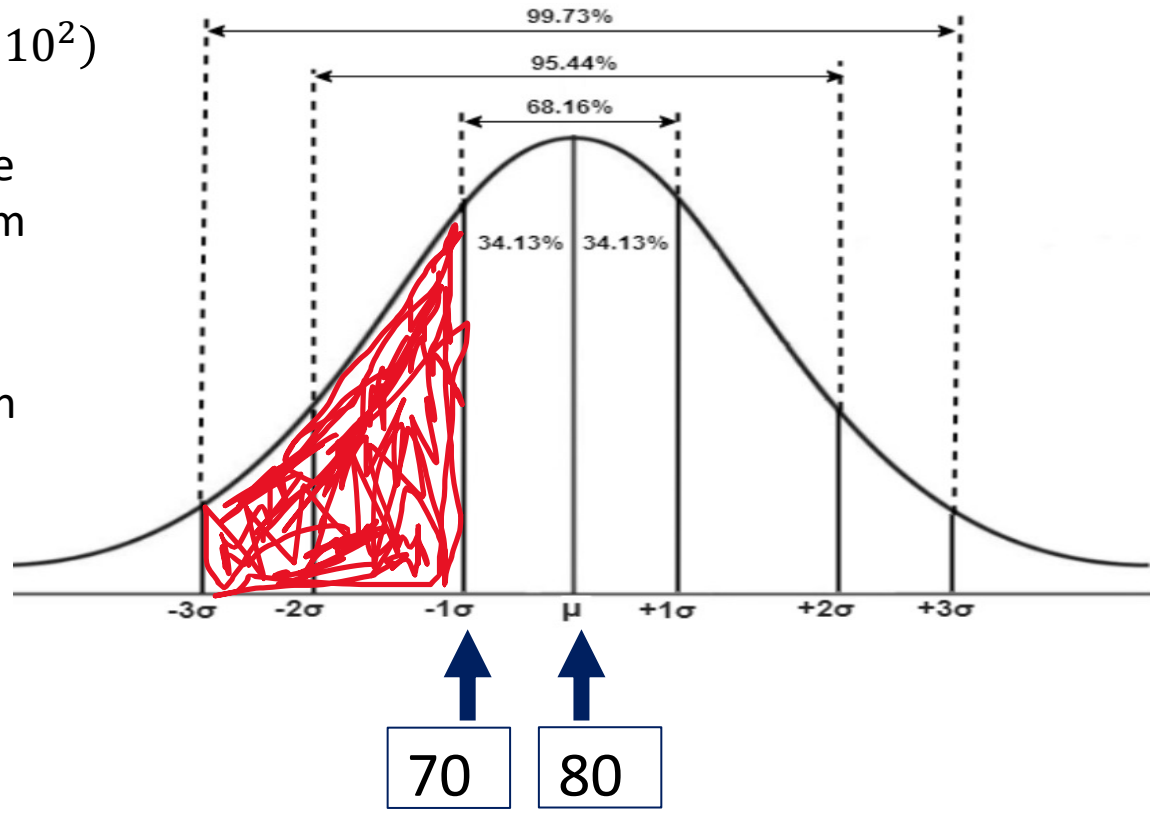
- Suppose that diastolic blood pressures (DBPs) from men aged 35-44 are normally distributed with a mean of 80 mmHg and a standard deviation of 10 mmHg (millimeters of mercury). About what is the probability that a random 35-44 years old has a DBP less than 70

  - 22%
  - 32%
  - 8%
  - 16%

Answer 2

we need to calculate the area under the normal distribution curve to the left of 70 mmHg.

- For normal distribution $N(80, 10^2)$
- Less than 70 lies outside of the one standard deviation(σ) from the mean in one side
- 68% of the normal distribution lies within one standard deviation



- We can use the Z-score formula to standardize the value of 70 mmHg:

$$Z = \frac{x - \mu}{\sigma}$$

$$x = 70\ mmHg$$
$$\mu = 80\ mmHg$$
$$\sigma = 10\ mmHg$$

$$Z = \frac{70 - 80}{10} = -1$$

Using a standard normal distribution table or a statistical calculator, the probability associated with a Z-score of -1 is approximately 0.1587.

$$P \approx 0.16 = 16\%$$

z	0.00
-3	0.0013
-2.9	0.0019
-2.8	0.0026
-2.7	0.0035
-2.6	0.0047
-2.5	0.0062
-2.4	0.0082
-2.3	0.0107
-2.2	0.0139
-2.1	0.0179
-2	0.0228
-1.9	0.0287
-1.8	0.0359
-1.7	0.0446
-1.6	0.0548
-1.5	0.0668
-1.4	0.0808
-1.3	0.0968
-1.2	0.1151
-1.1	0.1357
-1	0.1587
-0.9	0.1841

Quiz 5

- You flip a fair coin 5 times. About what is the probability of getting 4 or 5 heads?

  - 6%
  - 19%
  - 3%
  - 12%

Answer 5

- For each flipping, the probability for getting a head is a (Bernoulli distribution)

  - The total number of possible outcomes when flipping a fair coin 5 times is 2^5 = 32, as each flip has 2 possible outcomes (heads or tails) and we multiply them together for 5 flips.

  - To calculate the number of favorable outcomes, we need to consider two cases:

    - ✓ Getting 4 heads: There are 5 ways to choose which of the 5 flips will be heads, and the remaining flip will be tails. So, there are 5 favorable outcomes in this case.
    - ✓ Getting 5 heads: There is only 1 way to achieve this outcome.

$$P(4\ or\ 5\ heads) = \frac{favorable\ outcomes}{total\ number\ of\ possible\ outcomes}$$

- $P(4\ or\ 5\ heads) = \frac{1+5}{2^5} = 19\%$

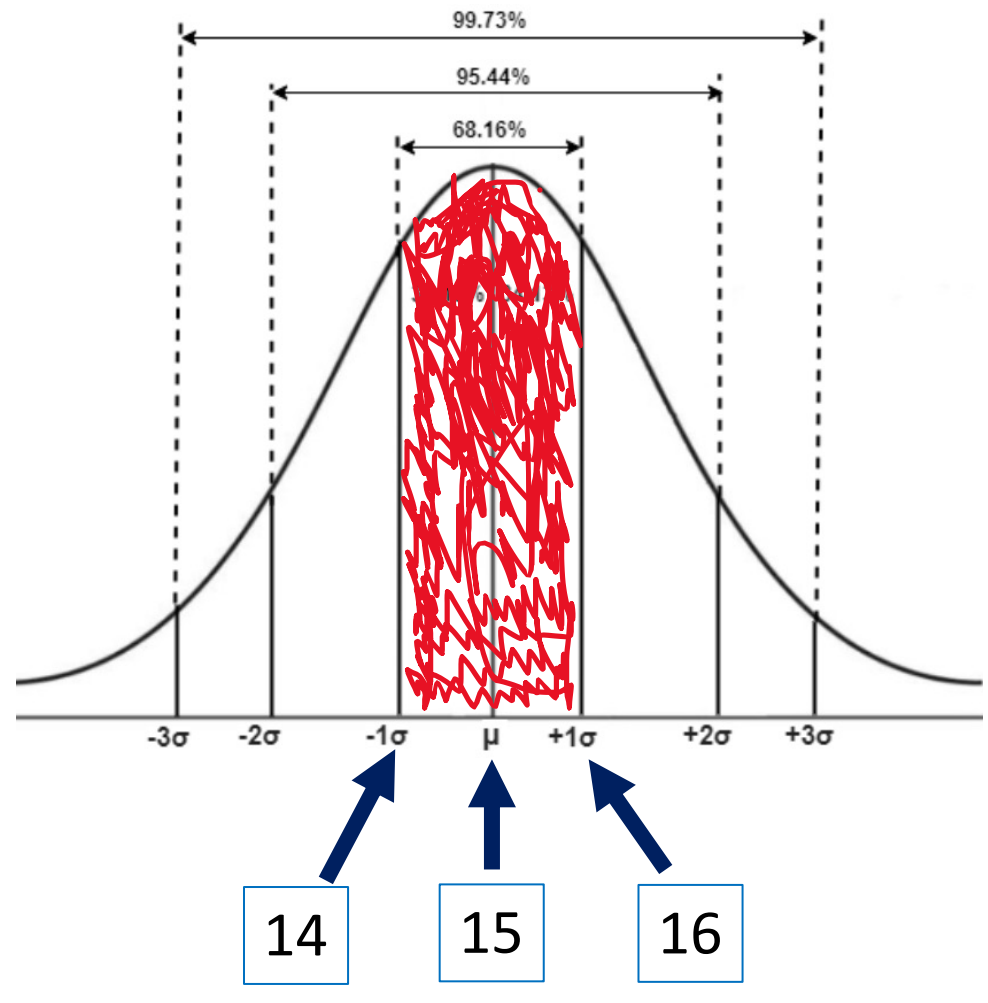
Quiz 6

- The respiratory disturbance index (RDI), a measure of sleep disturbance, for a specific population has a mean of 15 (sleep events per hour) and a standard deviation of 10. They are not normally distributed. Give your best estimate of the probability that a sample mean RDI of 100 people is between 14 and 16 events per hour.

  - 68%
  - 95%
  - 34%
  - 47.5%

Answer 6

- The standard error of the mean is $\frac{10}{\sqrt{100}} = 1$
- By CLT, sample mean $N\left(\mu, \frac{\sigma^2}{n}\right) = N(15, 1)$
- Thus between 14 and 16 is within one standard deviation of the mean of the distribution of the sample mean.
- Thus, it should be about 68%



Quiz 7

- Consider a standard uniform density. The mean for this density is 0.5 and the variance is 1/12. You sample 1000 observations from this distribution and take the sample mean, what value would you expect it to be near?

  - 0.75
  - 0.25
  - 0.10
  - 0.5

Answer 7

- By Law of Large Number, the sample mean will converge to the population mean if the sample size is large enough

- So, it should be near 0.5

Quiz 8

- The number of people showing up at a bus stop is assumed to be Poisson with a mean of 5 people per hour. You watch the bus stop for 3 hours. About what's the probability of viewing 10 of fewer people?
  - 0.08
  - 0.12
  - 0.03
  - 0.06

Answer 8

- $X \sim Pois(\lambda t) = Pois(15)$

- We want to calculate $P(X \leq 10)$which is the cumulative distribution function (CDF) of a Poisson distribution

$$P(X \leq 10) = \sum_{x=0}^{10} f(x) = \sum_{x=0}^{10} \frac{15^x e^{-15}}{x!}$$

- use R function

```
> ########################################## Quiz 8
> ppois(10, lambda = 15)
[1] 0.1184644
```