Module 5. Statistical Inferences: parameter estimation and confidence interval.

Overview:

In this module, we introduce the concept of statistical inference. Particularly, you will learn about how to estimate a parameter from data: with point estimation and interval estimation.

First, we shall introduce two common methods of deriving the point estimation (the best numerical guess of a parameter).

Then, we derive the sampling distribution of the mean and variance using the probability theory. These sampling distributions are used to derive interval estimation (confidence intervals). We discuss the common forms of confidence intervals.

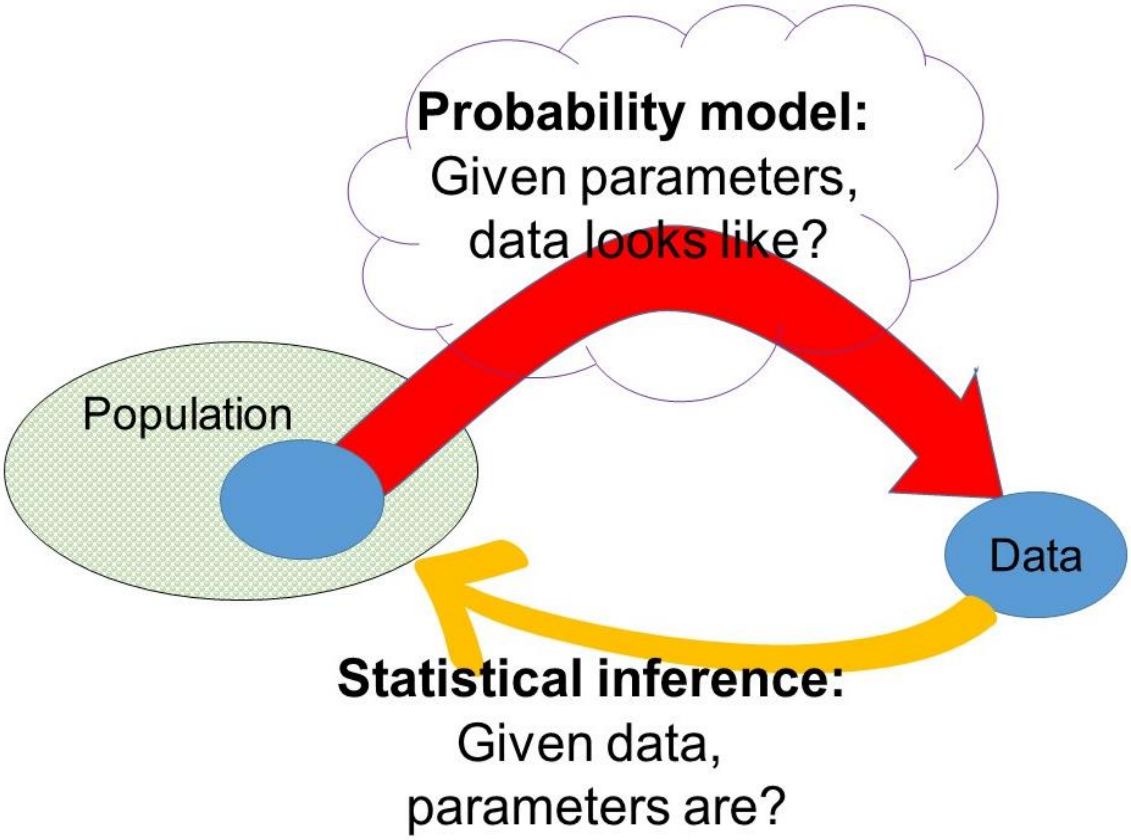
We also teach how to use Monte Carlo simulations to check the true coverage probability of confidence intervals. Finally, we introduce the nonparametric bootstrap confidence intervals.

Learning Objectives

- Use probability theory to **infer** values for the **population** given a sample data set
- Find **point estimations** by (a) **maximum likelihood** and (b) **Method of Moments**, both analytically and numerically
- Find the **sampling distribution** for sample mean and sample variance
- Calculate **confidence intervals** for mean and variance through sampling distribution (inlcuding Z-interval, t-interval, chi-square interval)
- Use **simulation** to check the coverage probability of confidence intervals
- Construct **Bootstrap** confidence intervals

Readings:
Seefeld & Linder’s book pages 217-239.
Krijnen’s book Pages 49-51.

This module introduce the basic concepts of statistical inference. Particularly, we consider the estimation of parameter, and the uncertainty measure for the estimation (i.e., confidence interval). These statistical inferences are derived from the corresponding probability theories.



Lesson 1. Two Methods of Parameter Estimation

We create parameter estimates using probability theory. In this lesson, we introduce two estimation methods. The first method involves maximizing the likelihood (probability) of observed data. The second method uses the sample moments (mean and variance)to estimate the corresponding theoretical moments (mean and variance).

Maximum Likelihood Estimation of Parameters: Probability Theory vs. Statistical Inference

Probability theory tells us how to calculate the probability (likelihood) of an observed data set given a parameter value. Statistical inference aims to find the parameter value given the observed data set. Let us look at an example.

Example 1: Maximum Likelihood for Poisson Observations

This sample uses three observations from Poisson distribution: 4, 7 and 2. Probability theory gives the density for $X_1 = 4$, $X_2 = 7$ and $X_3 = 2$ as $f(4,7,2;\lambda) = (\frac{\lambda^4}{4!}e^{-\lambda})(\frac{\lambda^7}{7!}e^{-\lambda})(\frac{\lambda^2}{2!}e^{-\lambda}) = \frac{\lambda^{13}}{4!7!2!}e^{-3\lambda}$, with the Poisson parameter λ. So if $\lambda = 1$, the pdf is $\frac{1^{13}}{4!7!2!}e^{-3} = 2.1\times10^{-7}$; if $\lambda = 3$, the pdf is $\frac{3^{13}}{4!7!2!}e^{-9} = 8.1\times10^{-4}$. We can use R to calculate these numbers as in Module 4.

```
> lik<-function(lam) prod(dpois(c(4,7,2), lambda=lam))
> c(lik(1), lik(3))
[1] 2.057997e-07 8.133064e-04
```

In the previous modules on probability, we consider the pdf $f(x_1, x_2, x_3; \lambda)$ mainly as a function with arguments x_1, x_2, x_3, and with a known λ value. Knowing λ, we use this pdf to see how likely we are to obtain various values of X_1, X_2 and X_3.

For the statistical inference, we identify $f(x_1, x_2, x_3; \lambda)$ as the likelihood function with an argument λ, and with the x_1, x_2, x_3 values known. The above calculation shows that, under $\lambda = 3$, the observed data occurs more likely than under $\lambda = 1$. A natural way of estimating the unknown λ is to maximize the likelihood function. This is called maximum likelihood estimation.

Here, we want to maximize the likelihood function $lik(\lambda) = f(4,7,2;\lambda) = \frac{\lambda^{13}}{4!7!2!}e^{-3\lambda}$.

A common mathematical trick is to equivalently maximize the logarithm of the likelihood instead: $l(\lambda) = \log[lik(\lambda)] = 13\log(\lambda) - 3\lambda - \log[4!7!2!]$.

Using calculus, we find the λ value for which the derivative equals zero: $l'(\lambda) = \frac{13}{\lambda} - 3 = 0$. That is, $\lambda = \frac{13}{3}$.

Thus, the maximum likelihood estimator (MLE) is $\hat{\lambda} = \frac{13}{3}$.

Maximum Likelihood for Poisson Observations: An Analytic Formula

Generally, for a random sample X_1, …, X_n from the Poisson(λ) distribution, the likelihood function is

$$lik(\lambda) = (\frac{\lambda^{X_1}}{X_1!}e^{-\lambda})...(\frac{\lambda^{X_n}}{X_n!}e^{-\lambda}) = \frac{\lambda^{\sum_{i=1}^{n} X_i}}{\prod_{i=1}^{n} X_i!}e^{-n\lambda}.$$

The log-likelihood function is

$$l(\lambda) = \log[lik(\lambda)] = (\sum_{i=1}^{n} X_i)\log\lambda - n\lambda - \log[\prod_{i=1}^{n} X_i!].$$

This is maximized by setting its derivative to zero,

$$l'(\lambda) = (\sum_{i=1}^{n} X_i)\frac{1}{\lambda} - n = 0.$$

Hence we get the analytic MLE (maximum likelihood estimator) formula as

$$\hat{\lambda} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Example 1 (continued):

Maximum Likelihood for Poisson Observations: Analytic Formula Versus Numerical Solution

For $X_1 = 4$, $X_2 = 7$ and $X_3 = 2$ from the Poisson(λ) distribution, applying the analytic MLE formula, we can quickly get $\hat{\lambda} = \bar{X} = \frac{1}{3}(4+7+2) = 4.3333$

However, even without an analytic formula, we can use R to numerically maximize the likelihood function.

```
> lik<-function(lam) prod(dpois(c(4,7,2), lambda=lam)) #likelihood function
> nlik<- function(lam) -lik(lam) #negative-likelihood function
> optim(par=1, nlik) #minimize nlik with starting parameter value=1
$par
[1] 4.333334
$value
[1] -0.001774955
$counts
function gradient
         48            NA
$convergence
[1] 0
$message
NULL
Warning message:
In optim(par = 1, nlik) :
   one-dimensional optimization by Nelder-Mead is unreliable:
use "Brent" or optimize() directly
```

We can see that this give us a numerical value for MLE $\hat{\lambda}$ =4.333334, agreeing with the answer using analytic formula.

In the R code here, we used the optim() routine which minimizes a function. Therefore, we minimized the negative likelihood function *nlik(λ)*, which is equivalent to maximizing the likelihood function *lik(λ)*.

More on Numerical Solution of MLE Poisson observations

Above we find the numerical value for maximum likelihood estimator (MLE) by using the likelihood function *lik(λ)* directly. As in the analytic derivation, this is equivalent to maximize the log-likelihood function *loglik(λ)*, or to minimize the negative log-likelihood function *nloglik(λ)*.

```
nloglik<- function(lam) -sum(log(dpois(c(4,7,2), lambda=lam)))
```

Using R to find numerical MLE by minimizing *nloglik(λ)*, we get

```
> optim(par=1, nloglik)$par #start at 1, minimize nloglik, ending parameter value
[1] 4.333594
```

This gives another approximate numerical solution $\hat{\lambda}$ =4.333594, very close to the previous answer.

In this case, using likelihood function directly gives more accurate answer. In some other cases, using the log-likelihood function can overcome some numerical stability issue.

Interested students can read the help file on the optimization method to find other ways of implementing MLE. Another way is to pass a negative control parameter inside optim() to do maximization instead of minimization. You can also use other routines such as optimize() or nlm().

Maximum Likelihood: Analytic Formula Versus Numerical Solution

We have seen in Example 1 that we can get the MLE either through analytic methods or numerical methods.

The analytic formula gives a more accurate MLE value. It also allows us to study uncertainty in the MLE through probability theory. We will use them to derive confidence intervals later in this module.

However, for more complicated models often encountered in bioinformatics applications, the analytic formulas may be too hard or impossible to derive. Numerical solutions will be very useful in those cases for getting the MLE. The confidence intervals for those MLE will have to come from more computationally intensive methods such as bootstrapping (which will be covered later in this module).

Example 2. Analytic MLE for Normal Observations

For a random sample $X_1, \ldots, X_n$ from the normal distribution $N(mean = \mu,\ sd = \sigma)$, the likelihood function is

$$lik(\mu,\sigma) = (\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(X_1-\mu)^2})\ldots(\frac{1}{\sqrt{2\pi}\sigma} e^{\frac{-1}{2\sigma^2}(X_n-\mu)^2}) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{\frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2} .$$

The log-likelihood function is

$$l(\mu,\sigma) = \log[lik(\mu,\sigma)] = \frac{-1}{2\sigma^2}\sum_{i=1}^{n}(X_i-\mu)^2 - n\log(\sigma) - n\log(\sqrt{2\pi}) .$$

Notice the parameters $(\mu,\ \sigma)$ are 2-dimensional. So the maximization would set both partial derivatives to be zero.

$$\frac{\partial}{\partial\mu} l(\mu,\sigma) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i-\mu) = \frac{1}{\sigma^2}[(\sum_{i=1}^{n} X_i) - n\mu] = 0 ;$$

$$\frac{\partial}{\partial\sigma} l(\mu,\sigma) = \frac{1}{\sigma^3}\sum_{i=1}^{n}(X_i-\mu)^2 - n(\frac{1}{\sigma}) = \frac{1}{\sigma^3}[\sum_{i=1}^{n}(X_i-\mu)^2 - n\sigma^2] = 0 .$$

Solving these two equations, we get the MLE for normal observations as

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X}, \quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\hat{\mu})^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} .$$

Example 3. Numeric MLE for Normal Observations

A random sample of size 5 from the distribution $N(mean = \mu,\ sd = \sigma)$ results in

-1.099 0.281 -0.726 -0.399 1.149

We can find the MLE numerically using the following code

```
obs<- c(-1.099, 0.281, -0.726, -0.399, 1.149) #save all observations in a vect
or
nloglik <- function(paras) { #negative-log-likelihood function (2 elements in paras)
       mu <- paras [1]          #call first parameter mu
      sigma <- paras [2]      #call second parameter sigma
      -sum(log(dnorm(obs, mean=mu, sd=sigma))) #negative log-likelihood
}
optim(par=c(2,2), nloglik)$par #maximize nloglik, start with mu=2, sigma=2
```

The answers are

```
[1] -0.1588497   0.7959305
```

Hence $\hat{\mu} = -0.1588497,\quad \hat{\sigma} = 0.7959305$. We can compare this with the answer from the analytic formula $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \bar{X},\quad \hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

```
> c(mean(obs), sqrt(sum((obs-mean(obs))^2)/5))
[1] -0.1588000   0.7958835
```

We can see that the analytic MLE and the numerical MLE agree to the third decimal space in this example.

Numeric MLE for Normal Distribution with Restricted Parameters

The normal distribution has two parameters μ and σ, so above we maximize a 2-dimensional function *nloglik(μ, σ)* to find the MLE. If we have a restricted normal distribution whose mean equals to its variance, then the distribution is reduced to a one-parameter distribution. That is, let $\theta = \mu = \sigma^2$. The normal distribution becomes $N(mean = \theta,\ sd = \sqrt{\theta})$.

If we again save the observations in a vector obs, then the negative log-likelihood function is defined as

```
nloglik <- function(theta) -sum(log(dnorm(obs, mean=theta, sd=sqrt(theta))))
```

and we can find MLE with

```
optim(par=1, nloglik)$par
```

In summary, to find MLE, we first write out the log-likelihood function, either analytically or in R. Then we either analytically maximize it or numerically maximize it in R.

Parameter Estimation by the Method of Moments (MoM)

Maximum Likelihood is not the only estimation method. Another commonly used method is to estimate the population quantities with the corresponding sample quantities. Particularly, the **method of moments** (MoM) uses sample moments to estimate the population moments.

The first moment of the population (that is, the mean or expectation) is μ = E(X). It is estimated by the sample mean $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

The second moment of the population (that is, the variance) is $Var(X) = E[(X - EX)^2]$. It is estimated by the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Correspondingly, the population standard deviation is estimated by the sample standard deviation $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

Example 4. MoM Estimator for Normal Distribution
For a random sample $X_1, \ldots, X_n$ from the normal distribution $N(mean = \mu,\ sd = \sigma)$, the population mean μ is estimated by $\hat{\mu} = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, and the population variance σ^2 is estimated by $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$. Hence σ is estimated by $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

Notice that the MoM (method of moments**)** estimator for μ is the same as the MLE. However, the MoM estimator for σ, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$, is different from the MLE $\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2}$.

The MLE mathematically is optimal in the asymptotic sense (that means for a big sample size n→∞). For a big sample size, the MLE and the MoM here will essentially be the same (n and n-1 will be very close). The divisor n-1 for the MoM estimator results in an unbiased estimator of σ^2 for any sample size.

Example 5. MoM Estimators for Chi-Square Distribution and Exponential Distribution

For a random sample X_1, …, X_n from the χ^2-distribution with degree of freedom m, the only parameter is m. We estimate it by matching the first moments of the population and of the sample. The population mean is m, the sample mean is $\bar{X}$. So the MoM estimator is $\hat{m} = \bar{X}$.

Let X_1, …, X_n be a random sample from the exponential distribution with parameter λ. We estimate λ by matching the first moments of the population and those of the sample. The population mean is 1/ λ, the sample mean is $\bar{X}$. So the MoM estimator is $\hat{\lambda} = \frac{1}{\bar{X}}$.

Summary

In this lesson we created parameter estimates using probability theory and estimation methods in general. The first method involves maximizing the likelihood (probability) of observed data. The second method uses the sample moments (mean and variance) to estimate the corresponding theoretical moments (mean and variance). In the next lesson we will work with sampling distributions to generate confidence intervals.

Lesson 2. Sampling Distributions.

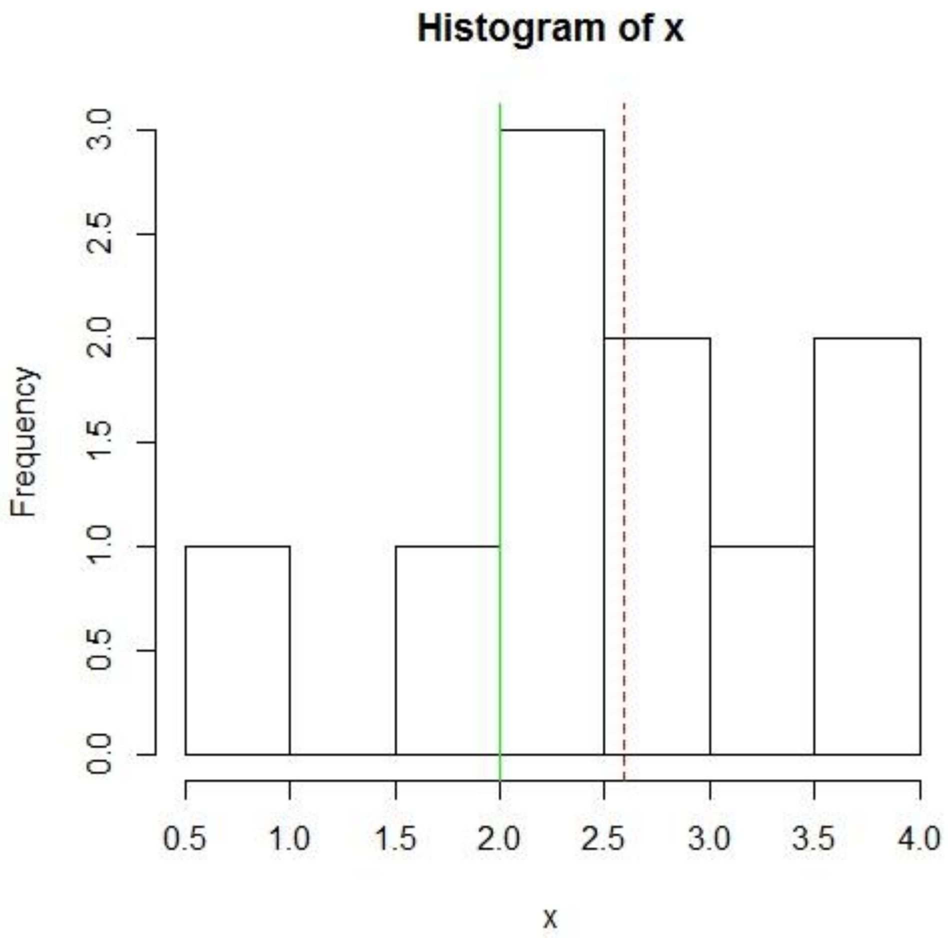
We have discussed the point estimator $\hat{\theta}$ for parameter θ. A very important part of statistical inference is to provide an accurate assessment of $\hat{\theta}$, in additional the simple point estimation of θ. The accuracy assessment is also provided through probability theory.

The observations $X_1,...,X_n$ are i.i.d. random variables in probability theory. Since the point estimator $\hat{\theta}$ is a function of $X_1,...,X_n$, $\hat{\theta}$ itself is also a random variable. The accuracy of the point estimator can be studied through its probability distribution. The probability distribution of $\hat{\theta}(X_1,...,X_n)$ changes with the sample size n, and is called as the "sampling distribution".

Some common sampling distributions are given in pages 217-226 of Seefeld and Linder's textbook. We will go over them next.

Sampling Distribution of the Sample Mean

The following graph shows the histogram of 10 observations from a normal distribution $N(\mu = 2, \sigma = 1)$. The green line indicate the population mean μ=2, while the red dotted line indicate the sample mean $\bar{X}$.



Given one data set, we would only see $\bar{X}$ (the red dotted line), the green solid line is unknown and is the object of our inference. To judge the error (difference between the two lines), we use the probability distribution of $\bar{X}$. This distribution is over many potential data sets, not this particular data set. The graph is generated by the following R code. If you ran the R code repeatedly, each run results in a different position for the red line, reflecting the sampling distribution of $\bar{X}$.

```
> x<-rnorm(10,mean=2,sd=1)
> hist(x)
> abline(v=2,col="green")
> abline(v=mean(x),col="red",lty=2)
```

Review of Properties of Linear Combination of Random Variables.

In probability theory, we have learned the following properties.

(1) For random variables $X_1, ..., X_n$, $E(a_1X_1 + ... + a_nX_n) = a_1EX_1 + ... + a_nEX_n$.

(2) For independent random variables $X_1, ..., X_n$,
$Var(a_1X_1 + ... + a_nX_n) = a_1^2Var(X_1) + ... + a_n^2Var(X_n)$.

(3) For normally distributed $X_1, ..., X_n$, $a_1X_1 + ... + a_nX_n$ is also normally distributed.

Using these properties on the sample mean, we get the following results:

$X_1, ..., X_n$ is a random sample of size n from a distribution with mean= μ and variance= σ^2. Then the sample mean statistic $\bar{X}$ has mean μ and variance $\frac{\sigma^2}{n}$.

Furthermore, if the distribution of X_1 is normal, then the sampling distribution of $\bar{X}$ is $N(mean = \mu, sd = \sigma / \sqrt{n})$.

Sampling Distribution of the Sample Mean(continued)

From last page, let X_1, …, X_n be a random sample from a normal distribution $N(mean = \mu, sd = \sigma)$. Then the sampling distribution of $\bar{X}$ is $N(mean = \mu, sd = \sigma / \sqrt{n})$.

The above result requires that all X_i's come from a normal distribution, which would restrict its usefulness in practice. Fortunately, we also have the central limit theorem (CLT) which states that this is the sampling distribution of $\bar{X}$ approximately in almost all cases.

Let X_1, …, X_n be a random sample from a distribution with mean μ and standard deviation σ<∞. Then for large sample size n, by CLT, the sampling distribution of $\bar{X}$ is approximately $N(mean = \mu, sd = \sigma / \sqrt{n})$.

Therefore, we can use the sampling distribution $N(mean = \mu, sd = \sigma / \sqrt{n})$ for inference always in practice (since it holds approximately for any distribution). We do need to be careful that a large sample size n is required if normality cannot be assumed.

Notice that the sampling distribution changes with n, as n affects the variance. For a large n, it is going to be very concentrated around the true mean μ.

Sampling Distribution of the Sample Mean: t-Distribution

Let X_1, …, X_n be a random sample from a normal distribution $N(mean = \mu, sd = \sigma)$. Then the sampling distribution of $\bar{X}$ is $N(mean = \mu, sd = \sigma / \sqrt{n})$. We can express this equivalently as

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

In practice, we generally do not know σ, and would like the sampling distribution in terms of the sample standard deviation

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$ instead of σ.

Theorem: With X_1, …, X_n being a random sample from $N(mean = \mu, sd = \sigma)$,

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$ follows the t-distribution with degree of freedoms n-1.

The quantity $s / \sqrt{n} = s.e.(\bar{X})$ is the estimated *standard error* of the sample mean $\bar{X}$.

Generally, if the normality assumption does not hold for X_i's, the t-distribution result still holds approximately for a large n, similar to the CLT.

Sampling Distribution of the Sample Variance: Chi-Square Distribution

To infer the variance, we can use the sampling distribution of the sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

Let X_1, …, X_n be a random sample from a normal distribution $N(mean = \mu, sd = \sigma)$, then the sampling distribution of $(n-1)\frac{s^2}{\sigma^2}$ is $\chi^2_{df} = n-1$.

Notice that this result does require normally distributed data X_i's.

Summary:

For a random sample X_1, …, X_n from $N(mean = \mu, sd = \sigma)$, we have the following three sampling distributions.

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1);$$

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{df\ = n-1};$$

$$(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{df\ = n-1} \ .$$

When the normality assumption does not hold, we still have approximately, for large sample size n,

$$\frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0,1) \text{ and}$$

$$\frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{df\ = n-1}.$$

These sampling distribution results will be used in the next lesson to derive confidence intervals.

Lesson 3. Confidence Intervals for Normal Mean.

Objectives

By the end of this lesson you will have had the opportunity to:

- Calculate confidence intervals for the population mean.
- Distinguish “confidence” from “probability”.

Overview

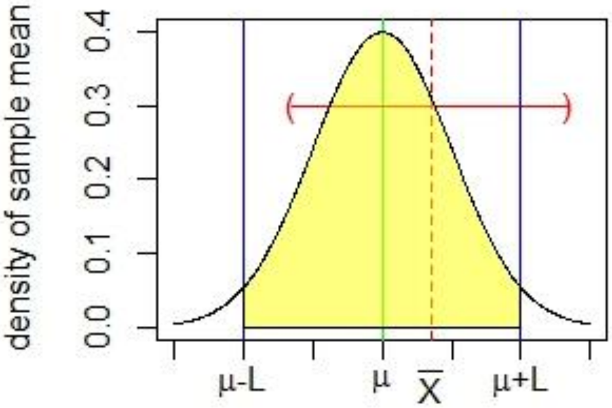
In this lesson, we introduce confidence intervals for the population. We distinguish the z-interval and t-interval by their different theoretical justifications. From those, we know when to use which interval formula.

The point estimator $\hat{\theta}$ only provides a single value as the best guess for parameter θ. We generally want to provide an interval estimate also for an accuracy assessment. That is, we use interval (θ_L, θ_U) for a range of possible values of θ.

This is formalized as a (1-α) *confidence interval* (CI): over repeated random samples in the long term, (1-α) proportion of CIs contain the true parameter θ. The formulas for the lower bound θ_L and upper bound θ_U can be derived from the sampling distribution.

Deriving Confidence Interval for Normal Mean.

The following graph shows the density curve of the sampling distribution of the sample mean $\bar{X}$, and the interval $(\theta_L = \bar{X} - L, \theta_U = \bar{X} + L)$ centered at $\bar{X}$ (position of the vertical red dotted line) for one data set.



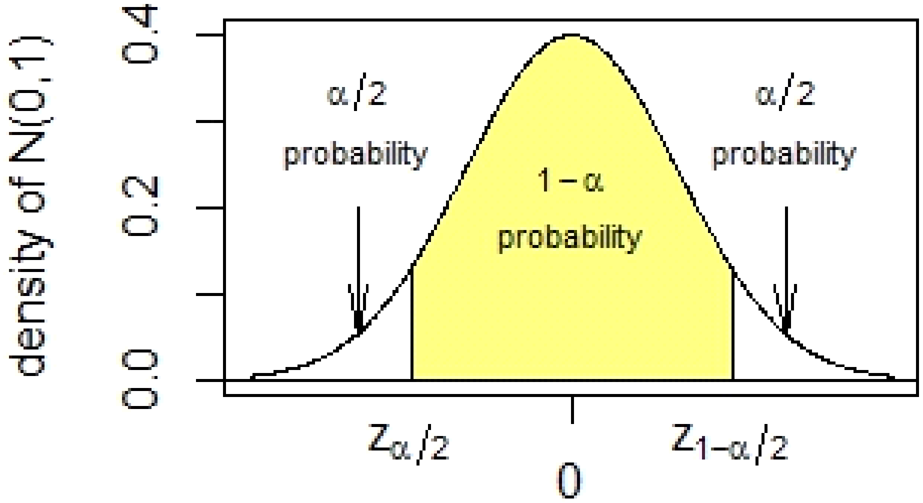
When $\bar{X} - L < \mu < \bar{X} + L$, the interval (shown as the line segment with parenthesis at the ends) contains true population mean μ (the vertical green line). From the graph, this is equivalent to $\mu - L < \bar{X} < \mu + L$ (when the red dotted line falls in between the two blue lines). Therefore the yellow area under the sampling distribution density curve represents the probability that the interval captures the true value over repeated sampling. We set it to 1-α to find a formula for L on the next section.

Confidence Interval for Normal Mean: z-Interval.

For the random sample of $X_1, \ldots, X_n$ □ $N(\mu,\sigma)$. The sampling distribution of $\bar{X}$ is $N(\mu,\sigma/\sqrt{n})$, i.e., $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ □ $N(0,1)$. We solve for the probability

1-α = P($\mu - L < \bar{X} < \mu + L$) = P($-\frac{L}{\sigma/\sqrt{n}} < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < \frac{L}{\sigma/\sqrt{n}}$).

Hence for 1-α confidence interval, we should let $-\frac{L}{\sigma/\sqrt{n}} = z_{\alpha/2}$ and $\frac{L}{\sigma/\sqrt{n}} = z_{1-\alpha/2}$ where $z_{\alpha/2}$ and $z_{1-\alpha/2}$ denote the α/2 and 1-α/2 quantiles of the standard normal distribution. They can be find in R using the qnorm() function.



Hence the 1-α CI is $(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}) = (\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$.

Confidence Interval for Normal Mean: z-Intervals.

For the random sample of X_1, …, X_n $\sim N(\mu,\sigma)$, the 1-α CI of μ is $(\bar{X}+z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\bar{X}+z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$.

Example 1.

For a random sample of X_1, …, X_{15} $\sim N(\mu,\sigma=2)$, if the sample mean $\bar{X}$ =3.12, find the 95% CI, the 90% CI and 80% CI.

Solution:

For the 95% CI, α=1-0.95=0.05. So $z_{\alpha/2}=z_{0.025}=-1.96$, $z_{1-\alpha/2}=z_{0.975}=1.96$.

Hence the 95% CI is $(\bar{X}+z_{0.025}\frac{\sigma}{\sqrt{15}},\bar{X}+z_{0.0975}\frac{\sigma}{\sqrt{15}})=(3.12-1.96\frac{2}{\sqrt{15}},3.12+1.96\frac{2}{\sqrt{15}})$ or (2.11, 4.13).

For the 90% CI, $z_{\alpha/2}=z_{0.05}=-1.645$, $z_{1-\alpha/2}=z_{0.95}=1.645$. The 90% CI is $(3.12-1.645\frac{2}{\sqrt{15}},3.12+1.645\frac{2}{\sqrt{15}})$ = (2.27, 3.97).

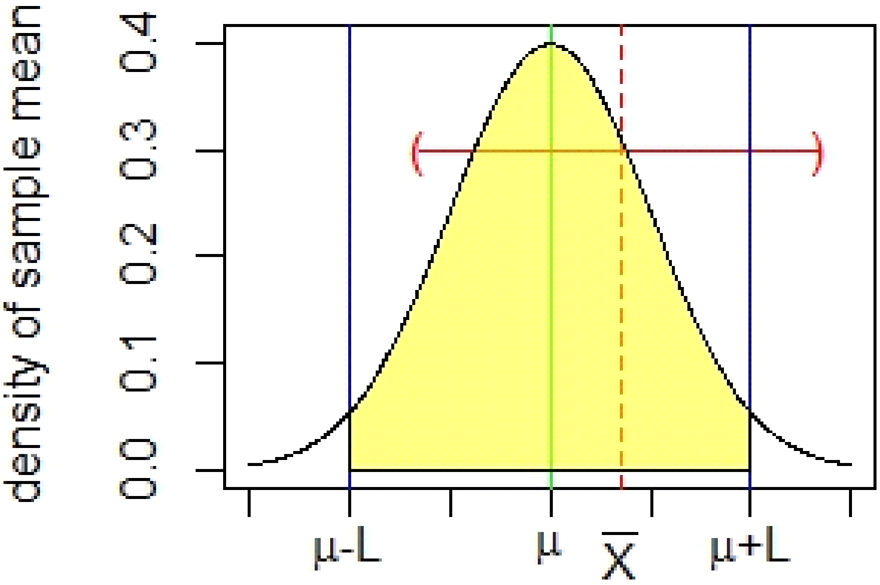
For the 80% CI, $z_{\alpha/2}=z_{0.1}=-1.28$, $z_{1-\alpha/2}=z_{0.9}=1.28$. The 80% CI is $(3.12-1.28\frac{2}{\sqrt{15}},3.12+1.28\frac{2}{\sqrt{15}})$ = (2.46, 3.78).

Confidence Intervals: Confidence versus Probability.

When we have a 95% CI of (2.11, 4.13), we say that we are "95% confident" that the true population mean μ is in the range between 2.11 and 4.13.

There is a natural tendency to claim that there is a "95% probability" that the true population mean μ is in the range between 2.11 and 4.13. However, this later statement is wrong. Why is it wrong? And why would we need the new term "confidence interval"?

Recall that we derived the formula of L in $\bar{X} \pm L$ from a probability statement. However, that probability statement is over many potential unseen data sets (the dotted red line will move around for different data sets, following the sampling distribution).



For the fixed data set, the dotted red line $\bar{X}$ is fixed. So the interval (2.11, 4.13) either captures the true green line or not, which we won't really know from the data set itself. The "95% confident" indicates a 95% probability about the moving random intervals of which (2.11, 4.13) is only one particular realization.

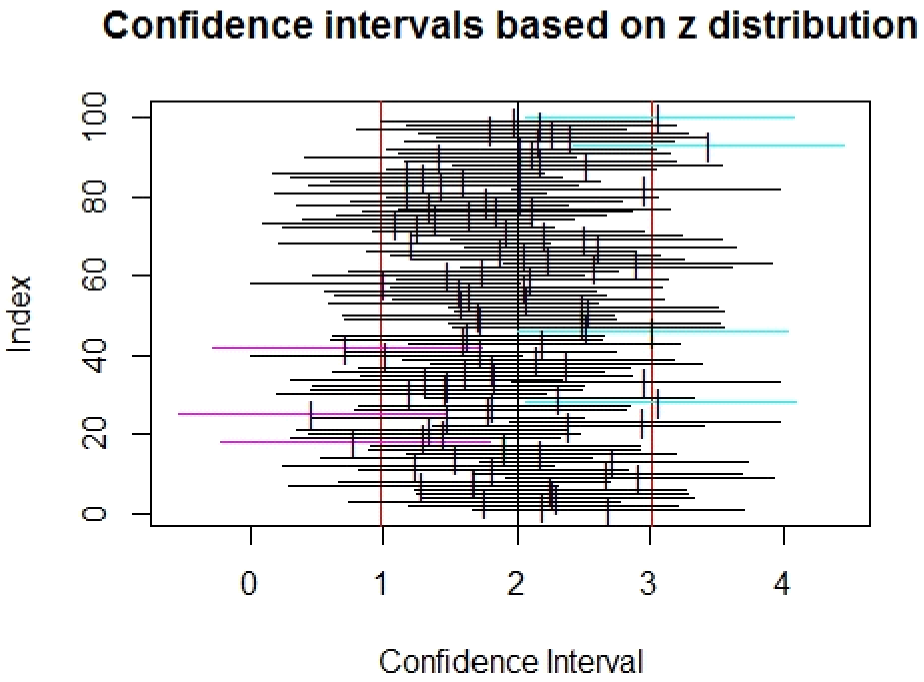
Confidence Intervals: Confidence versus Probability.

When we have a 95% CI of (2.11, 4.13), we say that we are “95% confident” that the true population mean μ is in the range between 2.11 and 4.13. This indicates a 95% probability of capturing the true μ by the moving random intervals of which (2.11, 4.13) is only one particular realization.

To illustrate this fact, we can run the ci.examp() from the R package *TeachingDemos*.

```
> library(TeachingDemos)
> ci.examp(mean.sim =2, sd = 2, n = 15, reps = 100, method = "z",
lower.conf=0.025, upper.conf=0.975)
```

Run this yourself. You can see that the 100 CIs will move around. The ones capturing true mean are in black (the blue and purple ones missed the true mean). Notice the number of CIs capturing the true mean in fact follow a binomial(size=100, prob=0.95) distribution, so you would get different numbers in different runs.



Confidence Interval for Normal Mean: Margin of Errors.

For the random sample of $X_1, \ldots, X_n \sim N(\mu, \sigma)$, the 1-α CI of μ is $\bar{X} \pm z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$.
This is an interval symmetric around the sample mean: in the form of $\bar{X} \pm L$.

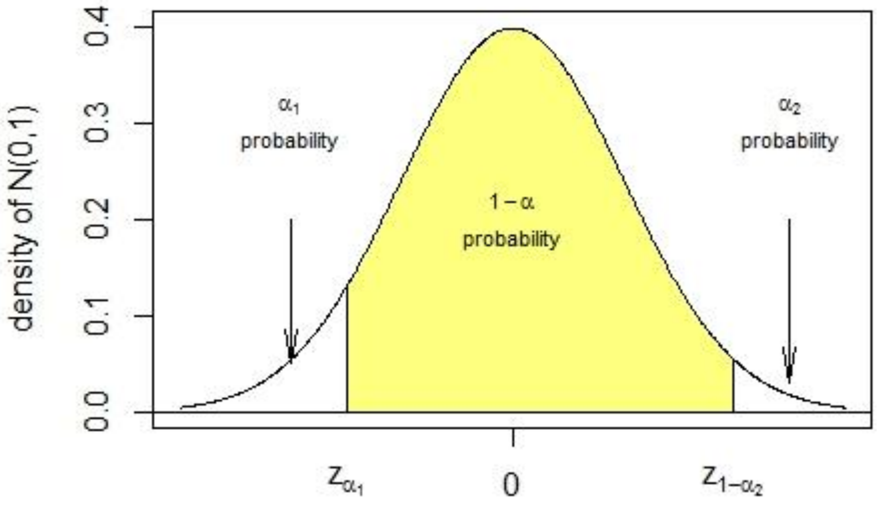
L, the half-length of the CI, is called the *margin of error*.
The margin of error has two components: the “confidence multiplier” -- $z_{1-\alpha/2}$ and $\frac{\sigma}{\sqrt{n}}$ -- the standard deviation of $\bar{X}$ (the point estimator of μ).

To get smaller margin of error, we can do the following: (1) reduces the confidence multiplier $z_{1-\alpha/2}$ (by using a lower confidence level 1-α) or (2) reduces the standard deviation $\frac{\sigma}{\sqrt{n}}$ of $\bar{X}$ (by increasing the sample size n).

This general structure applies to inference of quantities other than the population mean. That is, the margin of error comes from the two components: the confidence multiplier (decided by the sampling distribution) and the standard deviation of the estimator.

Asymmetric Confidence Interval for Normal Mean.

So far, we have considered only the symmetric 1-α CI of μ in the form of $\bar{X} \pm L$. That is, $(\bar{X} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$. We can use different quantiles to the left and to the right and get an asymmetric CI still with coverage probability of 1-α.



Let α_1+α_2=1, then $(\bar{X} + z_{\alpha_1}\frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\alpha_2}\frac{\sigma}{\sqrt{n}})$ is also a 1-α CI.

Particularly, we often are interested in only a lower (or upper) bound for μ. In such cases, the one-sided 1-α lower CI for μ is $(\bar{X} + z_{\alpha}\frac{\sigma}{\sqrt{n}}, \infty)$;

The one-sided 1-α upper CI for μ is $(-\infty, \bar{X} + z_{1-\alpha}\frac{\sigma}{\sqrt{n}})$.

Example 2.

For a random sample of X_1, …, X_{15} □ $N(\mu,\sigma=2)$, the sample mean $\bar{X}$ =3.12.

$z_{0.95}=1.645$, $z_{0.975}=1.96$.

The two-sided 95% CI is

$(3.12+z_{0.025}\frac{2}{\sqrt{15}},3.12+z_{0.0975}\frac{2}{\sqrt{15}})$ = (2.11, 4.13).

The one-sided lower 95% CI is

$(3.12+z_{0.05}\frac{2}{\sqrt{15}},\infty)=(3.12-1.645\frac{2}{\sqrt{15}},\infty)$ = (2.27, ∞).

The one-sided upper 95% CI is

$(-\infty,3.12+z_{0.95}\frac{2}{\sqrt{15}})=(-\infty,3.12+1.645\frac{2}{\sqrt{15}})$ = (-∞, 3.97).

Contrast this to the two-sided 90% CI, we get the following:

$(3.12-1.645\frac{2}{\sqrt{15}},3.12+1.645\frac{2}{\sqrt{15}})$ = (2.27, 3.97).

Confidence Interval for Normal Means: t-Intervals.

The z-interval formula involves the parameter value σ. In practice, σ is unknown. Therefore, a formula involving only the sample standard deviation s is needed.

For the random sample of $X_1, \ldots, X_n \sim N(\mu,\sigma)$, $\dfrac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{df = n-1}$. Using this sampling distribution to solve for the probability

1-α = P($\mu - L < \bar{X} < \mu + L$) = P($-\dfrac{L}{s/\sqrt{n}} < \dfrac{\bar{X}-\mu}{s/\sqrt{n}} < \dfrac{L}{s/\sqrt{n}}$), we get $-\dfrac{L}{s/\sqrt{n}} = t_{\alpha/2,n-1}$ and $\dfrac{L}{s/\sqrt{n}} = t_{1-\alpha/2,n-1}$ where $t_{\alpha/2,n-1}$ and $t_{1-\alpha/2,n-1}$ denote the α/2 and 1-α/2 quantiles of the t-distribution distribution with degree of freedom n-1. They can be calculated using the R function qt(x, df=n-1).

Hence the two-sided 1-α CI for μ is $(\bar{X} + t_{\alpha/2,n-1}\dfrac{s}{\sqrt{n}}, \bar{X} + t_{1-\alpha/2,n-1}\dfrac{s}{\sqrt{n}}) = \bar{X} \pm t_{1-\alpha/2,n-1}\dfrac{s}{\sqrt{n}}$.

Similarly, the one-sided CI for μ is $(\bar{X} + t_{\alpha,n-1}\dfrac{s}{\sqrt{n}}, \infty)$ or $(-\infty, \bar{X} + t_{1-\alpha,n-1}\dfrac{s}{\sqrt{n}})$.

Example 3.

For a random sample of X_1, …, X_9 □ $N(\mu,\sigma)$, the sample mean and sample standard deviation are $\bar{X}$ =4.7, s=18.2.

Hence the two-sided 80% CI is

$(\bar{X}+t_{0.1,8}\frac{s}{\sqrt{9}}, \bar{X}+t_{0.9,8}\frac{s}{\sqrt{9}}) = (4.7-1.39(\frac{18.2}{\sqrt{9}}), 4.7+1.39(\frac{18.2}{\sqrt{9}}))$ = (-3.8, 13.2).

Notice that (-3.8, ∞) and (-∞, 13.2) are one-sided 90% CIs.

The one-sided 80% CIs are $(\bar{X}+t_{0.2,8}\frac{s}{\sqrt{9}}, \infty) = (4.7-0.89(\frac{18.2}{\sqrt{9}}), \infty)$ = (-0.69, ∞) and $(-\infty, 4.7+0.89(\frac{18.2}{\sqrt{9}}))$ = (-∞, 10.1).

Example 4.

Generally, given a data set, we can use R to calculate the sample mean, the sample standard deviation and the corresponding CI (t-interval).

We illustrate this on the Gdf5 gene expression data in the Golub et al. (1999) data set. (See Example 1 on page 48 of Krijnen's textbook.) The patients are classified into two groups: ALL(Acute lymphocytic leukemia) and AML(acute myeloid leukemia). We calculate the CIs for the mean Gdf5 gene expression on these two groups separately.

The data set is loaded from the *multtest* package from Bioconductor. If you have not done so, install it on your machine using the following R commands (page 2-3 of Krijnen's textbook).

```
source("http://www.bioconductor.org/biocLite.R")
biocLite("multtest")
```

Then calculate the CIs by R

```
> data(golub, package = "multtest") #load "golub" data in package "multtest"
> gol.fac <- factor(golub.cl,levels=0:1,labels=c("ALL","AML")) #get a factor
variable, label values 0/1 to ALL/AML
> x <- golub[2058, gol.fac=="ALL"] #take the 2058th row, and columns for
 ALL patients
> n<-length(x)   #sample size n
> ci.all <- mean(x)+qt(c(0.025,0.957),df=n-1)*sd(x)/sqrt(n) #95% t-interval
> x <- golub[2058,gol.fac=="AML"] #take the 2058th row, and columns for
AML patients
> n<-length(x)
> ci.aml <- mean(x)+qt(c(0.025,0.975),df=n-1)*sd(x)/sqrt(n)
> print(ci.all) #print the 95% CI for ALL patients (stored in vector ci.all)
[1] -0.1024562 0.1025636
> print(ci.aml)
[1] -0.06450576  0.22532213
```

The 95% Cis in ALL group is (-0.10, 0.10) and in the AML group is (-0.06, 0.23). The significant overlap in these CIs indicates no difference between mean gene expressions in the two groups. This indicates that the Gdf5 gene is not related to the leukemia.

Summary:

For a random sample X_1, …, X_n from $N(mean = \mu, sd = \sigma)$, we have two types of confidence intervals for μ: z-interval and t-interval.

The z-interval for μ: σ is known. Standard deviation $sd(\bar{X}) = \frac{\sigma}{\sqrt{n}}$.

Two-sided 1-α CI: $(\bar{X} + z_{\alpha/2} \cdot sd(\bar{X}), \bar{X} + z_{1-\alpha/2} \cdot sd(\bar{X}))$;

One-sided 1-α CI: $(\bar{X} + z_{\alpha} \cdot sd(\bar{X}), \infty)$ and $(-\infty, \bar{X} + z_{1-\alpha} \cdot sd(\bar{X}))$.

The t-interval for μ: σ is unknown. Standard error $se(\bar{X}) = \frac{s}{\sqrt{n}}$.

Two-sided 1-α CI: $(\bar{X} + t_{\alpha/2,n-1} \cdot se(\bar{X}), \bar{X} + t_{1-\alpha/2,n-1} \cdot se(\bar{X}))$;

One-sided 1-α CI: $(\bar{X} + t_{\alpha,n-1} \cdot se(\bar{X}), \infty)$ and $(-\infty, \bar{X} + t_{1-\alpha,n-1} \cdot se(\bar{X}))$.

In practice, we always use t-intervals for data analysis since σ is unknown.

The z-interval, in practice, is only used for sample-size determination when we have prior estimation of σ. The sample-size determination can be found in many introductory statistics textbooks. We do not discuss that usage in this course.

Lesson 4. Confidence Intervals for Mean and Variances.

For a random sample X_1, …, X_n from $N(mean = \mu, sd = \sigma)$, we have derived the confidence intervals for μ, from the two sampling distributions $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\dfrac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{df = n-1}$.

We can derive confidence interval formulas for more quantities using other sampling distributions.

Confidence Interval for Variance.

For the random sample of $X_1, \ldots, X_n \sim N(\mu, \sigma)$, we have the sampling distribution $(n-1)\frac{s^2}{\sigma^2} \sim \chi^2_{df} = n-1$.

Hence a 1-α CI for population variance σ^2 is $(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}})$.

Notice that this CI is not centered at the point estimator s^2. This is because the chi-squared distribution is skewed, so that we do not want a symmetric confidence interval.

The one-side 1-α CI for σ^2 is $(\frac{(n-1)s^2}{\chi^2_{1-\alpha,n-1}}, \infty)$ or $(0, \frac{(n-1)s^2}{\chi^2_{\alpha,n-1}})$.

The confidence intervals for population standard derivation σ can be found corresponding. The two-sided 1-α CI for σ is $(\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{1-\alpha/2,n-1}}}, \frac{s\sqrt{n-1}}{\sqrt{\chi^2_{\alpha/2,n-1}}})$. The one-sided 1-α CI for σ is $(\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{1-\alpha,n-1}}}, \infty)$ or $(0, \frac{s\sqrt{n-1}}{\sqrt{\chi^2_{\alpha,n-1}}})$.

Example 1

For a random sample of $X_1, \ldots, X_{16}$ □ $N(\mu,\sigma)$, $\bar{X}$ =4.7, s=2.2.

Therefore a 95% CI for population variance σ^2 is

$$\left(\frac{15(2.2)^2}{\chi^2_{0.975,15}}, \frac{15(2.2)^2}{\chi^2_{0.025,15}}\right) = \left(\frac{15(2.2)^2}{27.5}, \frac{15(2.2)^2}{6.26}\right) \ = (2.64,\ 11.6).$$

A 95% CI for population standard deviation σ is $\left(\frac{\sqrt{15}(2.2)}{\sqrt{27.5}}, \frac{\sqrt{15}(2.2)}{\sqrt{6.26}}\right)$ = (1.6, 3.4).

A 95% one-sided CI for σ is $\left(\frac{\sqrt{15}(2.2)}{\sqrt{\chi^2_{0.95,15}}}, \infty\right) = \left(\frac{\sqrt{15}(2.2)}{\sqrt{25.0}}, \infty\right)$ = (1.7, ∞).

Confidence Interval for Population Mean.

Until now, we have derived confidence intervals assuming that the data uses $\sim N(\mu,\sigma)$. What if we cannot assume that the data comes from a normal distribution? Fortunately we recall that the sampling distributions $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$ and $\frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{df\,=n-1}$ still hold approximately for large sample size n.

Hence we use the same z-intervals and t-intervals formulas as before on all data sets when n is large.

As a rule of thumb, generally the approximation is good for n≥30. (You can try out the CLT applet at https://adamding.shinyapps.io/CLTadamding/)

Comment: For n≥30, the t-multiplier $t_{\alpha,n-1}$ is very close to the z-multiplier z_α. Generally the z-intervals are used in such cases for convenience (we do not look up different quantiles for different degree of freedoms n-1). However, nowadays we do all the calculations in R, as it takes no more effort to find $t_{\alpha,n-1}$ than to find z_α in R. *Hence, it is recommended that you always use the t-interval for real data analysis.*

Confidence Interval for Poisson Mean.

For a random sample X_1, …, X_n from the Poisson(λ) distribution, the point estimator (MLE, MoM) for λ is $\hat{\lambda} = \bar{X}$. What is the interval estimation?

For large n, we can use the z-interval $(\bar{X} + z_{\alpha/2} \cdot se(\bar{X}), \bar{X} + z_{1-\alpha/2} \cdot se(\bar{X}))$ as a 1-α CI for λ.

Since Poisson variance is λ, we use $\hat{\lambda} = \bar{X}$ instead of sample variance $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$ to estimate the population variance. Hence $se(\bar{X}) = \sqrt{\frac{\bar{X}}{n}}$.

Therefore the two-sided 1-α CI for λ is $(\bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}}{n}}, \bar{X} + z_{1-\alpha/2}\sqrt{\frac{\bar{X}}{n}})$.

The one-sided 1-α CI for λ is $(\bar{X} + z_{\alpha}\sqrt{\frac{\bar{X}}{n}}, \infty)$ or $(-\infty, \bar{X} + z_{1-\alpha}\sqrt{\frac{\bar{X}}{n}})$.

Example 2

A random sample of 50 observations from a Poisson distribution have the sample mean $\hat{\lambda} = \bar{X} = 8.42$.

The a 90% CI for λ is

$(\bar{X} + z_{0.05}\sqrt{\frac{\bar{X}}{50}}, \bar{X} + z_{0.95}\sqrt{\frac{\bar{X}}{50}}) = (8.42 - 1.645\sqrt{\frac{8.42}{50}}, 8.42 + 1.645\sqrt{\frac{8.42}{50}})$ = (7.745, 9.095).

A 92% CI for λ is $(\bar{X} + z_{0.04}\sqrt{\frac{\bar{X}}{50}}, \bar{X} + z_{0.96}\sqrt{\frac{\bar{X}}{50}}) = (8.42 - 1.75\sqrt{\frac{8.42}{50}}, 8.42 + 1.75\sqrt{\frac{8.42}{50}})$ = (7.70, 9.14).

Confidence Interval for Binomial Proportions.

For a random sample X from the binomial(size=n, prob=p) distribution, the point estimator (MLE, MoM) for p is $\hat{p} = \frac{X}{n}$. What is the interval estimation?

Since the binomial random variable is the sum of n independent Bernoulli random variables, the central limit theorem applies here. Hence for large samples, we use the z-interval $(\hat{p} + z_{\alpha/2} \cdot se(\hat{p}), \hat{p} + z_{1-\alpha/2} \cdot se(\hat{p}))$ as a 1-α CI for λ.

Since $\mathrm{var}(\hat{p}) = \mathrm{var}(\frac{X}{n}) = \frac{\mathrm{var}(X)}{n^2} = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}$, $se(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$.

Therefore the two-sided 1-α CI for p is $(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$.

The one-sided 1-α CI for λ is $(\hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \infty)$ or $(-\infty, \hat{p} + z_{1-\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$.

As a rule of thumb, this CI formula works well when np≥10 and n(1-p)≥10.

This CI is often called the Wald interval, named after the statistician Abraham Wald.

Example 3
An observation from the Binomial(70, p) distribution is X=52.

Then the point estimation is the sample proportion $\hat{p}=\frac{52}{70}=0.743$.

A 95% CI for p is $(0.743+z_{0.025}\sqrt{\frac{0.743(1-0.743)}{70}},0.743+z_{0.975}\sqrt{\frac{0.743(1-0.743)}{70}})$ which can be calculated as (0.640, 0.845) using R command

```
>  p<-52/70
>  p+qnorm(c(0.025,0.975))*sqrt(p*(1-p)/70)
[1] 0.6404715 0.8452428
```

A 90% one-sided lower CI for p is $(0.743+z_{0.1}\sqrt{\frac{0.743(1-0.743)}{70}},1]$ = (0.676, 1].

That is, we are 90% confident that the true proportion p is great than 67.6%.

Summary:

For a random sample $X_1, \ldots, X_n$ from normal distribution, the confidence intervals for population mean and population standard deviation are of the form $(\bar{X} + t_{\alpha/2,n-1} \cdot se(\bar{X}), \bar{X} + t_{1-\alpha/2,n-1} \cdot se(\bar{X}))$ and $(\frac{s\sqrt{n-1}}{\sqrt{\chi^2_{1-\alpha/2,n-1}}}, \frac{s\sqrt{n-1}}{\sqrt{\chi^2_{\alpha/2,n-1}}})$ respectively.

For non-normal data, asymptotically (large **n**), we also use z-intervals for population mean $(\bar{X} + z_{\alpha/2} \cdot se(\bar{X}), \bar{X} + z_{1-\alpha/2} \cdot se(\bar{X}))$. The standard error $se(\bar{X})$ is based on the parametric distribution.

For the binomial(n, p) data, the proportion p is estimated by $\hat{p} = \frac{X}{n}$ with standard error $se(\hat{p}) = \frac{\hat{p}(1-\hat{p})}{n}$. The 1-α CI for the binomial proportion is therefore $(\hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + z_{1-\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}})$.

When n is small and data are non-normal, there are no general CI formulas for the population mean. We will revisit this in the next module on hypothesis testing.

Lesson 5. Bootstrap Confidence Intervals.

Objectives

By the end of this lesson you will have had the opportunity to:

- Construct Bootstrap confidence intervals
- Use simulation to check the coverage probability of confidence intervals

Overview

In this lesson, we will continue our study of confidence intervals. Particularly, we will introduce Bootstrapping confidence intervals. We will define them and calculate them using R. We will also use Monte Carlo simulation to compare the coverage probability of the bootstrap interval with previous t-interval.

Bootstrap Confidence Intervals.

For non-normal data, sometimes a very large n is needed for the central limit theorem (CLT) to take effect. In such cases, the nonparametric bootstrap confidence intervals often perform better than the CLT-based z-intervals and t-intervals.

Bootstrapping refers to the process of resampling from observed data to produce many pseudo data sets. These pseudo data sets are used to estimate the variability of a point estimator such as the sample mean. The same bootstrap procedure works for any other point estimator such as sample standard deviation, sample median, etc. Thus the bootstrap is very useful in producing confidence intervals when the theoretical sampling distributions are complicated or unknown.

Nonparametric Bootstrap Confidence Intervals

For a random sample X_1, …, X_n, we describe the nonparametric bootstrap CI for the sample mean $\hat{\theta}(X_1,...,X_n) = \bar{X}$. The same bootstrap procedure works for any other statistic $\hat{\theta}(X_1,...,X_n)$.

Randomly sampling with replacement from (X_1, …, X_n), we get a pseudo-data set ($X^*_{1,1}$, …, $X^*_{1,n}$) and can calculate pseudo-statistic $\widehat{\theta}^*_1 = \hat{\theta}(X^*_{1,1},...,X^*_{1,n}) = \frac{1}{n}\sum_{i=1}^{n} X^*_{1,i}$. Repeat this B times, we get pseudo-statistics $\widehat{\theta}^*_b = \hat{\theta}(X^*_{b,1},...,X^*_{b,n})$ for b=1,…,B. (See figure 13-11 on page 237 of Seefeld & Linder's textbook).

Then the 1-α CI for θ (the population mean here) is ($\theta^*_{\alpha/2}$, $\theta^*_{1-\alpha/2}$) where $\theta^*_{\alpha/2}$ and $\theta^*_{1-\alpha/2}$ are the α/2 and 1-α/2 quantiles of the bootstrapped pseudo-statistics $(\widehat{\theta}^*_1,...,\widehat{\theta}^*_B)$.

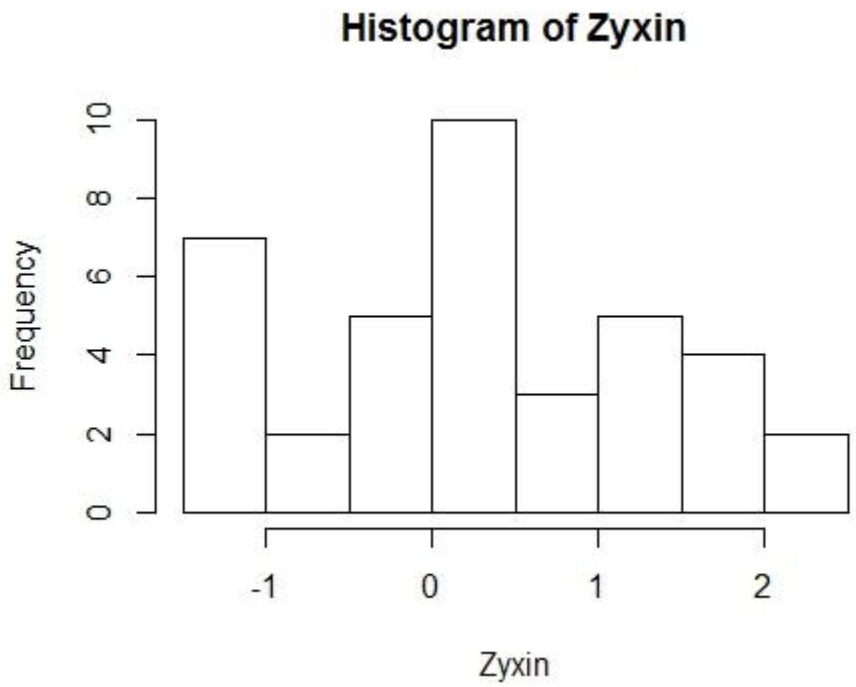
For the number of bootstraps B, the larger the better. In practice, B=1000 generally suffices for producing CIs.

Example 5.5.1 CI for the Mean Expression of the Zyxin Gene

In this example, we use the Golub et al. (1999) data set (available in the *multtest* package). The Zyxin gene can be found in row 2124 by searching the gene names in golub.gnames.

```
> data(golub, package = "multtest")
> grep("Zyxin",golub.gnames[,2])
[1] 2124
```

We load the gene data and resample B=1000 times, then find the quantiles to get the nonparametric bootstrap 95% CI for the population mean (-0.07, 0.59).
In this case, the bootstrap CI is close to the t-interval (-0.10, 0.60) since the data is approximately bell-shaped.



Example(continued) using R.

The bootstrap CI was calculated by the following R script.

```
Zyxin<-golub[2124,]
n<-length(Zyxin)
nboot<-1000
boot.xbar <- rep(NA, nboot)
for (i in 1:nboot) {
   data.star <- Zyxin[sample(1:n,replace=TRUE)]
   boot.xbar[i]<-mean(data.star)
}
quantile(boot.xbar,c(0.025,0.975))
```

The t-interval is calculated by

```
mean(Zyxin)+qt(c(0.025,0.975),df=n-1)*sd(Zyxin)/sqrt(n)
```

Monte Carlo Simulation and Bootstrap

Strictly speaking, bootstrapping is a kind of Monte Carlo procedure. However, most of the Monte Carlo simulation repeatedly samples from a theoretical distribution. Bootstrap repeatedly samples from the empirical distribution, essentially treating the observed data set as a theoretical distribution.

The bootstrap procedure can be proven to provide valid statistical inference asymptotically (when sample size n is large). This is similar to the t-interval for population mean. When n is small, however, the bootstrap CI cannot be proven valid.

When to Use the Bootstrap CI

The bootstrap procedure used here assumes the data is a random sample. That is, X_1, …, X_n are i.i.d. random variables. (This is an important assumption often not stated explicitly.) If the X_i's are dependent, more efforts are needed to provide valid bootstrap procedures.

The main advantage of the nonparametric bootstrap CI is its simplicity. It works asymptotically for any statistic, without the need to derive the sampling distributions which may be theoretically difficult.

Next, we illustrate the advantage of bootstrap CI on the sample median.

Confidence Intervals for the Sample Median

It is straightforward to produce the nonparametric bootstrap CI for the sample median. Calculate the sample median $\widehat{\theta}_b^* = median(X_{b,1}^*, ..., X_{b,n}^*)$ on each bootstrapped pseudo-data set, and then finding the interval ($\theta_{\alpha/2}^*$, $\theta_{1-\alpha/2}^*$) from the quantiles of $(\widehat{\theta}_1^*, ..., \widehat{\theta}_B^*)$.

Deriving a CI for the median of the sampling distribution is much harder. The distribution of the sample median (which is not covered in most statistics textbooks) asymptotically is approximately normal with mean m and standard deviation $\frac{1}{2f(m)\sqrt{n}}$. Here m denotes the population median, and f(m) is the density function value at m. Therefore, the 1-α CI for m is $(\hat{m} + z_{\alpha/2}\frac{1}{2f(\hat{m})\sqrt{n}}, \hat{m} + z_{1-\alpha/2}\frac{1}{2f(\hat{m})\sqrt{n}})$.

The disadvantages of the second approach include the following:

(1) We need to find the sampling distribution of the sample median.

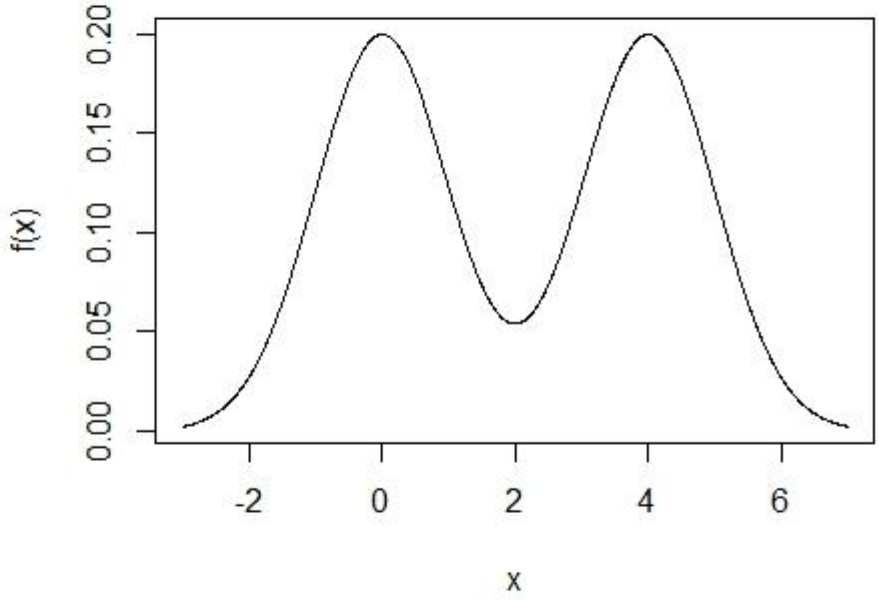
(2) A very large sample size is required for asymptotic approximation (much larger n compared to that required for CLT of the sample mean approximation).

(3) The density f(m) needs to be estimated, which is a very hard statistical problem.

Example 2: CI for the Sample Median

IN this example, we study the real coverage probabilities of three confidence intervals for sample median in a non-normal data example. The first CI is the bootstrap CI ($\theta^*_{\alpha/2}$, $\theta^*_{1-\alpha/2}$). The second is the z-interval $(\hat{m}+z_{\alpha/2}\frac{1}{2f(\hat{m})\sqrt{n}}, \hat{m}+z_{1-\alpha/2}\frac{1}{2f(\hat{m})\sqrt{n}})$ where the f(m) is estimated from normal distribution $f(\hat{m})=\frac{1}{\sqrt{2\pi}s}e^{0}=\frac{1}{\sqrt{2\pi}s}$. The third CI is also the z-interval where f(m) is estimated by the default density estimation procedure density() in R.

A Monte Carlo simulation is conducted with data sets of size n=30 generated from a mixture normal distribution, whose density is plotted below.



The Monte Carlo coverage probabilities are 94.8%, 65.6% and 76.6% for the three CIs respectively. Here, the bootstrap CIs perform (94.8%) very similar to the nominal 95% level. The z-intervals performs badly (65.6% and 76.6%) due to incorrect density estimation and slow asymptotic convergence (n=30 is not big enough).

The simulation code is in CImedian.r **(add link to the file)**. You can run it yourself. The numbers will change for each run, but the general pattern will be the same: bootstrap CI performs much better.

Example 3: CI for the Median Expression of Zyxin Gene

The bootstrap CI for the median expression can be calculated with a method similar to the one used for the mean expression, with a quick modification of the R code. (The red colored word is the only change.)

```
Zyxin<-golub[2124,]
n<-length(Zyxin)
nboot<-1000
boot.xbar <- rep(NA, nboot)
for (i in 1:nboot) {
    data.star <- Zyxin[sample(1:n,replace=TRUE)]
    boot.xbar[i]<-median(data.star)
}
quantile(boot.xbar,c(0.025,0.975))
```

Run it and we get the 95% for the median expression of the Zyxin gene as (-0.22, 0.57).

Module Summary:

In this module, we learned two basic statistical inferences of the parameter: the point estimator and the confidence interval. You should learned how do we derive these inference methods from probability theory, and know the standard methods for several basic cases.

1. You should know two methods of point estimation: maximum likelihood estimator (MLE) and the method of moments (MoM). You should know how to find them analytically and numerically from the data.
2. You should know how to derive confidence intervals from the sampling distributions and how to check the correctness of CI formulas using Monte Carlo simulation.
3. You should know how to calculate the CI for population mean and population variance for data from normal distributions. You should also be able to calculate CI for other quantities related to population mean: such as the binomial proportion.
4. You should know how to calculate a bootstrap CI for any statistic.