Statistical Data Analysis

Assist. Prof. Dr. Zeyneb KURT

(Slides have been prepared by
Prof. Dr. Nizamettin AYDIN,
updated by Zeyneb KURT)

zeyneb@yildiz.edu.tr
http://avesis.yildiz.edu.tr/zeyneb/

Estimation (cont’d)

Parameter Estimation

- The aim of statistics is to make inferences about a population based on information contained in a sample.
- Most inferential problems can be formulated as an inference about one or more parameters of a population (e.g. mean μ, the median M, the standard deviation σ, and a proportion π).
- Methods for making inferences about parameters fall into one of two categories:
  - estimate the value of the population parameter of interest
  - test a hypothesis about the value of the parameter

Parameter Estimation

- These two methods of statistical inference involve different procedures, and they answer two different questions about the parameter.
  - In estimating a population parameter, we are answering the question
    - "What is the value of the population parameter?"
  - In testing a hypothesis, we are seeking an answer to the question
    - "Does the population parameter satisfy a specified condition? "
- We will discuss statistical methods for parameter estimation (we will guess the unknown value of a parameter e.g., population mean, using the observed data.
- For this, we will use an estimator, which is a statistic.
  - A statistic is a function of the observed data only.
- Sometimes we only provide a single value as our estimate.
  - This is called point estimation.
    - Point estimates do not reflect our uncertainty when estimating a parameter.
    - We always remain uncertain regarding the true value of the parameter when we estimate it using a sample from the population.
- To address this issue, we can present our estimates in terms of a range of possible values.
  - This is called interval estimation.

Convention

- We use $X_1, X_2, \ldots, X_n$ to denote n possible values of X obtained from a sample randomly selected from the population.
- We treat $X_1, X_2, \ldots, X_n$ themselves as n random variables because their values can change depending on which n individuals we sample.
- We assume the samples are independent and identically distributed (IID).
- While theoretically we can have many different samples of size n, we usually have only one such sample in practice.
- We use $x_1, x_2, \ldots, x_n$ as the specific set of values we have observed in our sample.
- That is, x_1 is the observed value for X_1, x_2 is the observed value X_2, and so forth.

Point estimation - Population Mean

- Sometimes we only provide a single value as our estimate.
  - This is called point estimation.
    - We use $\hat{\mu}$ and $\hat{\sigma}^2$ to denote the point estimates for and μ and σ^2.
- For a population of size N, μ is calculated as.

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

- Given n observed values, $X_1, X_2, \ldots, X_n$, from the population, we can estimate the population mean μ with the sample mean:

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

  - In this case, we say that $\bar{X}$ is an estimator for μ.
- As our sample (the n representative members from the population) changes, the value of this estimator (sample mean) can also change.

Point estimation - Population Mean

- We usually have only one sample of size n from the population, x_2, … , x_n .
- Therefore, we only have one value for $\bar{X}$, which we denote

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

where x_i is the ith observed value of X in our sample, and $\bar{x}$ is the observed value of $\bar{X}$.

- As an example,
  - {consider the study* to estimate the population mean for body temperature among healthy people. From a sample of $n = 148$ people, they estimated the unknown population mean with the sample mean $\hat{\mu} = \bar{x} = 98.25$. This estimate is lower than the commonly believed value of 98.6°F.}
    - [The sample size for this study was relatively small. We would expect that as the sample size increases, our point estimate based on the sample mean would become closer to the true population mean.]

*Mackowiak, P.A., Wasserman, S.S., Levine, M.M.: A critical appraisal of 98.6°F, the upper limit of the normal body temperature, and other legacies of Carl Reinhold AugustWunderlich. JAMA 268, 1578–1580 (1992)

Point estimation - Population Variance

- The population variance is the average of squared deviations of each observation x_i from the population mean μ and denoted as σ^2

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Given n randomly sampled values $X_1, X_2, \ldots, X_n$ from the population and their corresponding sample mean $\bar{X}$, we can estimate the variance as :

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

- However, this estimator tends to underestimate the population variance.

Point estimation - Population Variance

- To address this issue, a more commonly used estimator for σ^2 is the sample variance:

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

  - This is the sum of squared deviations from the sample mean divided by $n-1$ instead of n.
    - Dividing by $n-1$ instead of n increases the value of the estimator by a small amount, which is enough to avoid underestimation associated with the more natural estimator.
- Therefore, the sample variance is the usual estimator of the population variance.
  - Likewise, the sample standard deviation S, $(\sqrt{S^2})$, is our estimator of the population standard deviation σ.
- We regard the estimator S^2 as a random variable since it changes as we change the sample.

Point estimation - Population Variance

- However, in practice, we usually have one set of observed values, $x_1, x_2, \ldots, x_n$, and therefore, only one value for S^2, denoted as s^2:

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- For binary random variables with 0 and 1 values, we can show that the population variance σ^2 is equal to $\mu(1-\mu)$, where μ is the population mean (proportion).
  - (See the Bernoulli distribution)
- Therefore, after we estimate the population mean μ using the sample mean (proportion) $\bar{x} = p$, we can use it to estimate the population variance instead of estimating σ^2 separately:

$$s^2 = p(1-p)$$

Sampling distribution

- The value of estimators discussed so far (and all estimators in general) depend on the specific sample selected from the population.
- If we repeat our sampling, we are likely to obtain a different value for an estimator.
  - Therefore, we regard the estimators themselves as random variables.
  - As a result, we can talk about their probability distribution.
- Probability distributions for estimators are called sampling distributions.
- Here, we are mainly interested in the sampling distribution of the sample mean $\bar{X}$.
  - For binary random variables, this is the same as the sample proportion.

Sampling distribution

- We start by assuming that the random variable of interest, X, has a normal $N(\mu, \sigma^2)$ distribution.
- Further, we assume that the population variance σ^2 is known, so the only parameter we want to estimate is μ.
- We need to find the sampling distribution of $\bar{X}$ under these assumptions.
  - {As a running example, consider the random variable $X \sim N(125, 15^2)$ representing systolic blood pressure, whose population mean $\mu = 125$ is unknown to us, but we know the population variance $\sigma^2 = 15^2$.
    - The population standard deviation is $\sigma = 15$.}

Sampling distribution

- Suppose that we take a sample of size $n = 2$ from the population.
- The corresponding values obtained from this sample are denoted as X_1 and X_2, which assumed to be identically distributed and independent.
- We write this as

$$X_1, X_2 \sim N(\mu, \sigma^2)$$

- Because they are independent and identically distributed (IID), their sum is also normally distributed,

$$X_1 + X_2 \sim N(\mu + \mu, \sigma^2 + \sigma^2) = N(2\mu, 2\sigma^2)$$

- This can be generalized as

$$X_1 + X_2 + \cdots + X_2 \sim N(n\mu, n\sigma^2) \quad \text{or} \quad \sum_{i=1}^{n} X_\mathrm{i} \sim N(n\mu, n\sigma^2)$$

Sampling distribution

- If $\sum_{i-1}^{n} X_\mathrm{i}$ divided by n, the sample mean is obtained:

$$\bar{X} = \frac{\sum_{i-1}^{n} X_\mathrm{i}}{n}$$

- When we multiply a random variable by a constant (here, $1/n$), its mean is multiplied by that constant, and its variance is multiplied by the square of that constant.
- So, if we multiply $\sum_{i-1}^{n} X_\mathrm{i}$ by $1/n$ to obtain the sample mean $\bar{X}$, the mean becomes $n\mu/n = \mu$, and the variance becomes $n\sigma^2/n^2 = \sigma^2/n$.
- In this case,

$$\bar{X} \sim N(\mu, \sigma^2/n)$$

where n is the sample size.

Sampling distribution

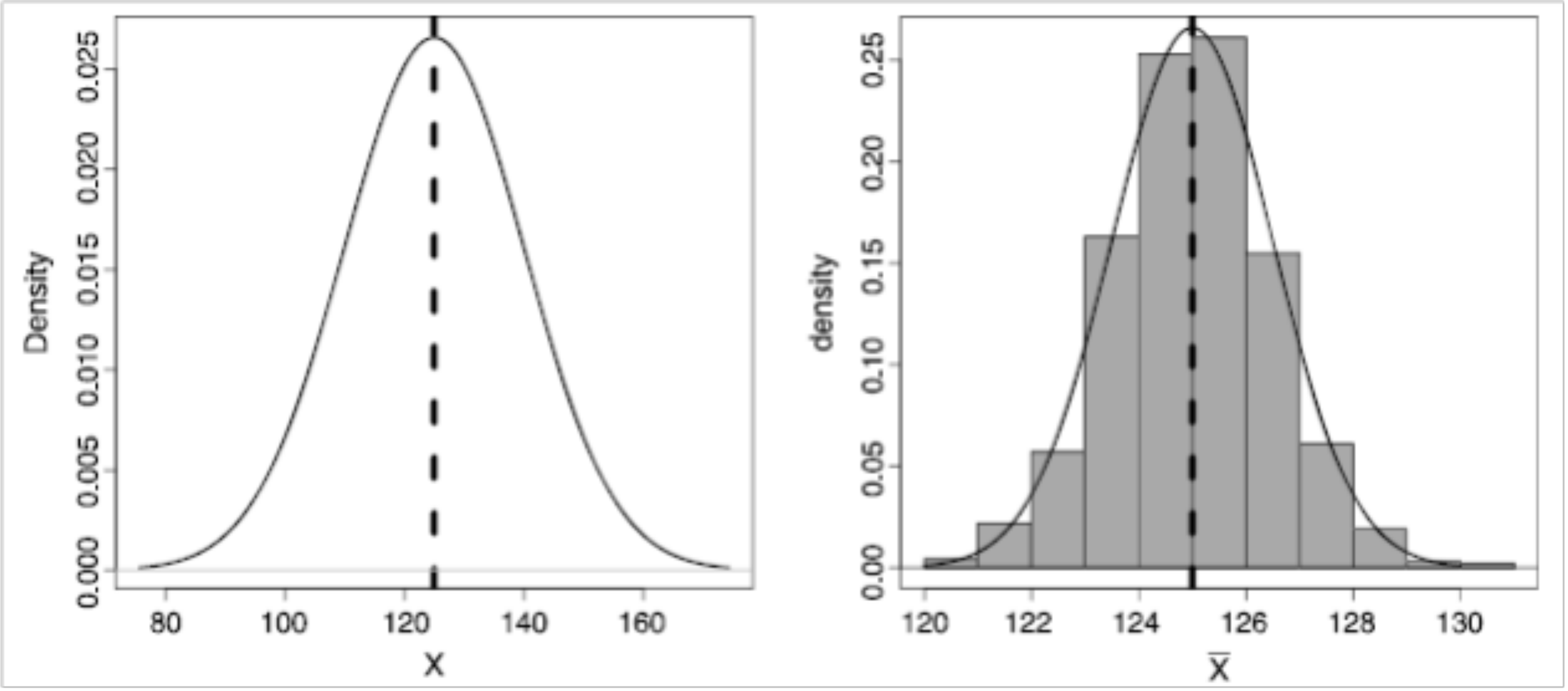
- The standard deviation of $\bar{X}$ can be obtained by taking the square root of its variance:

$$\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$$

- The standard deviation of the sampling distribution in this case reflects the extent of the variability of the sample mean as an estimator for the population mean.
  - For the above blood pressure example, if we take a sample of size $n = 100$ from the population and use $X_1, X_2, \ldots, X_{100}$ to denote the 100 possible values obtained from this sample, we have

$$X_1, X_2, \ldots, X_{100} \sim N(125, 15^2)$$
$$\bar{X} \sim N(125, 15^2/100)$$

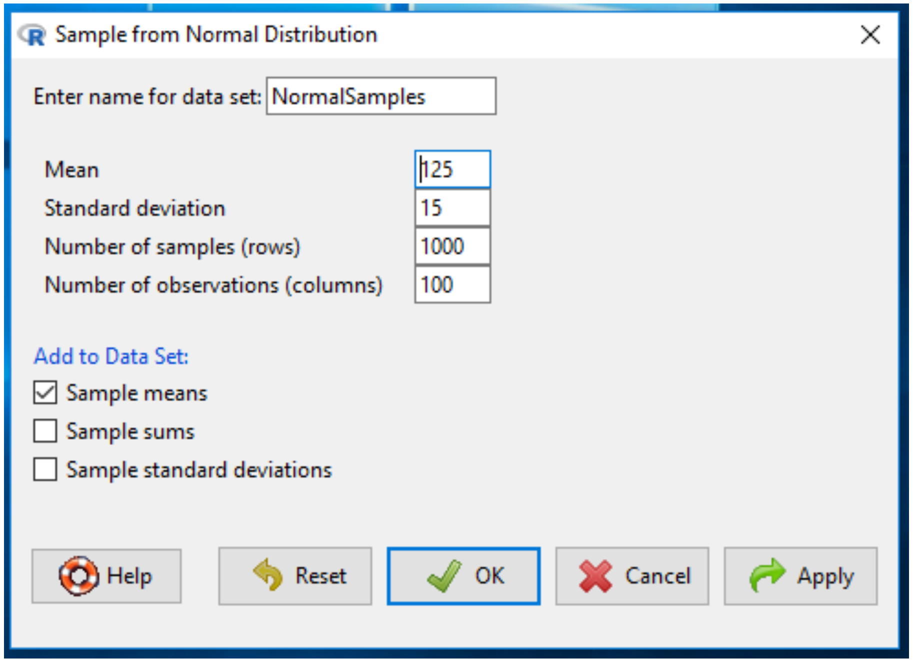
Sampling distribution



- *Left panel*: The (unknown) theoretical distribution of blood pressure, $X \sim N(125, 15)$.
- *Right panel*: The density curve for the sampling distribution $\bar{X} \sim N(125, 15^2/100)$ along with the histogram of 1000 sample means.
    - The distribution of sample means is centered on the population mean (shown with the a *vertical line*), but its variance is much less than that of blood pressure itself.
        - Note the different scales on the x-axis

Sampling distribution

- We can simulate the procedure using R-Commander.
  - Click *Distributions* → *Continuous distributions* → *Normal distribution* → *Sample from normal distribution*.
  - Then enter 125 for the *mean*, 15 for *standard deviation*.



Enter name for data set: NormalSamples

Mean: 125
Standard deviation: 15
Number of samples (rows): 1000
Number of observations (columns): 100

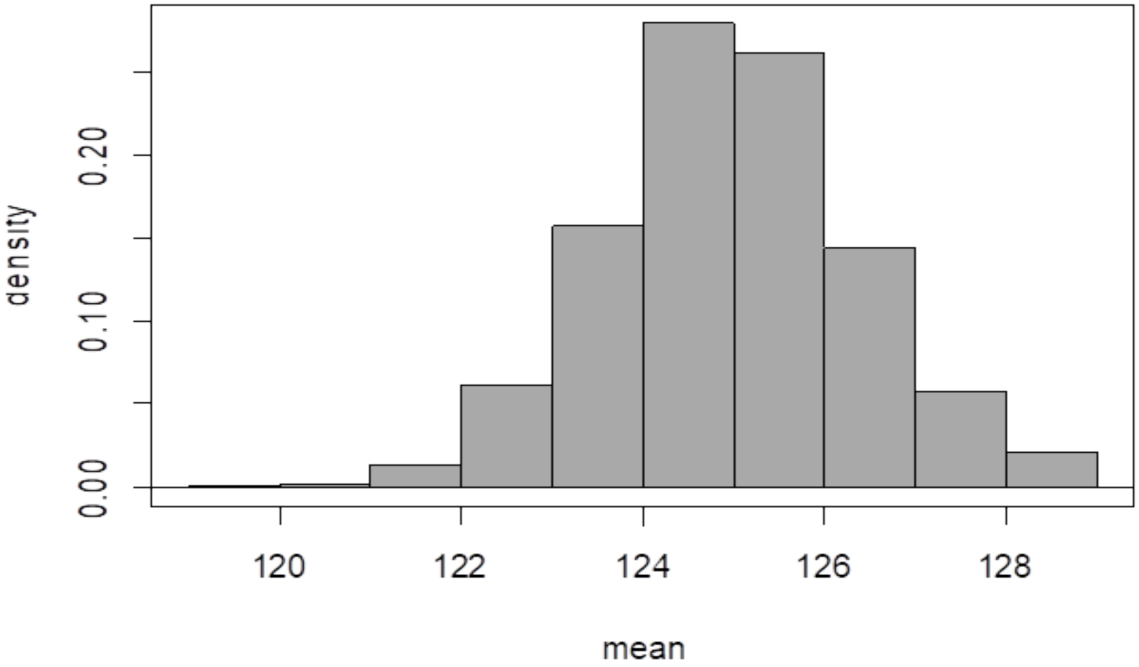
Add to Data Set:

☑ Sample means
☐ Sample sums
☐ Sample standard deviations

  - Set the *Number of samples (rows)* to 1000 and the *Number of observations (columns)* to 100, as in the figure.
    - This creates 1000 different samples, where the size of each sample is $n = 100$.
  - Keep the option *Sample means* checked;
    - this will store the sample means in a variable called mean.

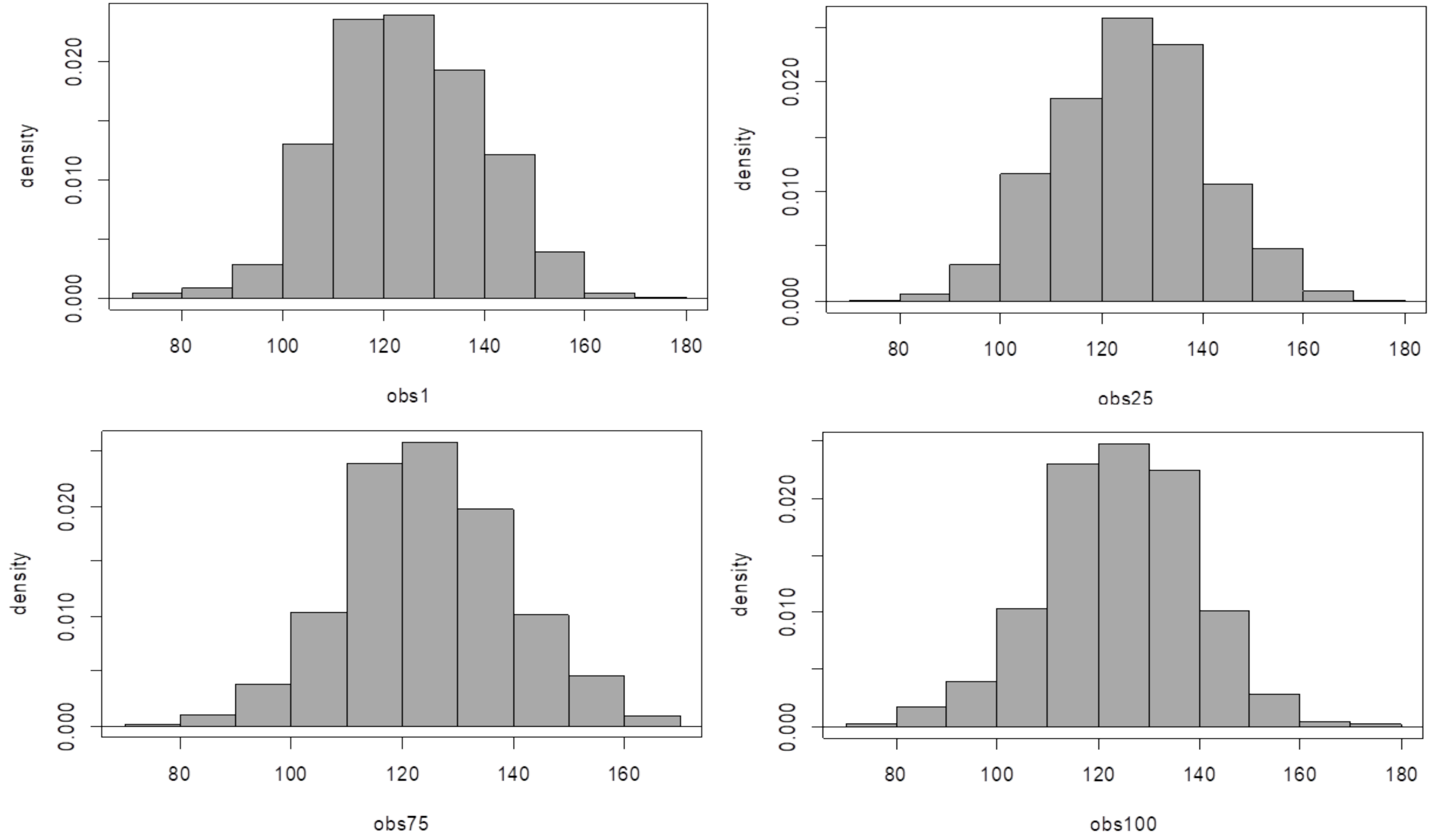
Sampling distribution

- We can now plot the histogram of the 1000 sample means. (NormalSamples should be the active data set.)



- Click *Graphs* → *Histograms*; choose mean as the *Variable* in Data tab and check *Densities* in Options tab.

Sampling distribution



Confidence Intervals for the Population Mean

- It is common to express our point estimate along with its standard deviation to show how much the estimate could vary if different members of population were selected as our sample.
- Alternatively, we can use the point estimate and its standard deviation to express our estimate as a range (interval) of possible values for the unknown parameter.

Confidence Intervals for the Population Mean

- Consider the estimation of the population mean μ in the systolic blood pressure example
- We know that $\bar{X} \sim N(\mu, \sigma^2/n)$
- Since the sampling distribution is normal,
  - the 68–95–99.7% rule applies.
- Therefore, approximately 95% of the values of $\bar{X}$ fall within the 2 standard deviations of the mean.

Confidence Intervals for the Population Mean

- Suppose that $\sigma^2 = 15^2$ and sample size is $n = 100$.
- So, the SD of $\bar{X}$ is $\sigma/\sqrt{n} = 1.5$
- Following the 68–95–99.7% rule, with 0.95 probability, the value of $\bar{X}$ is within 2 SDs from its mean, μ,

$$\mu - 2\times1.5 \leq \bar{X} \leq \mu + 2\times1.5$$

- In other words, with probability 0.95,

$$\mu - 3 \leq \bar{X} \leq \mu + 3$$

Confidence Intervals for the Population Mean

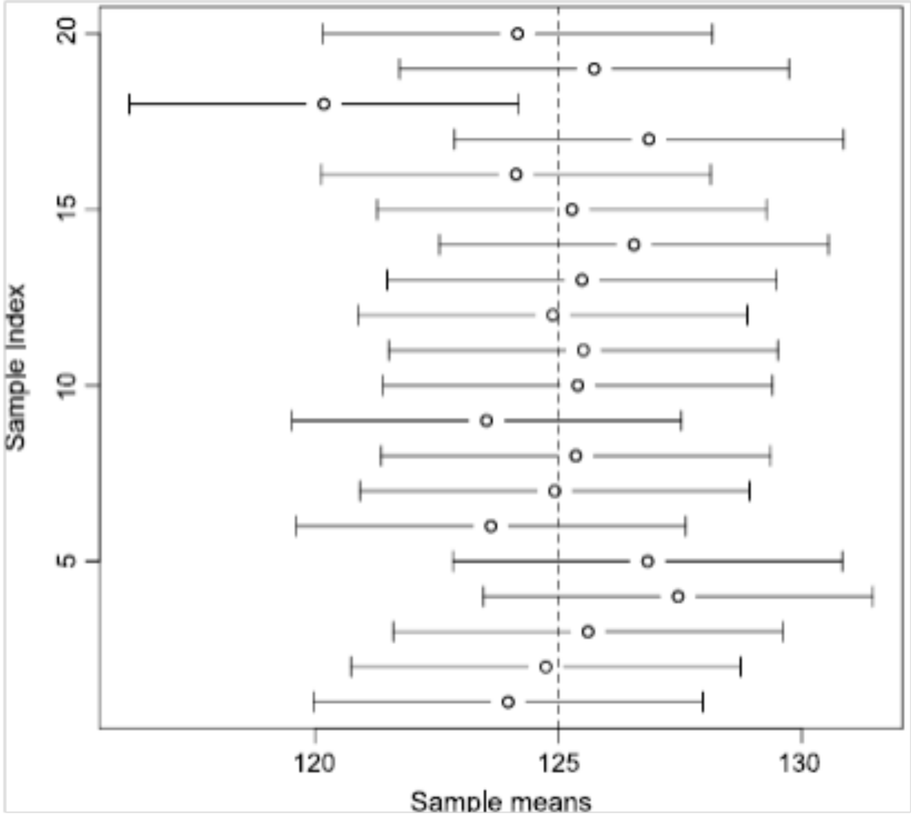
- We are, however, interested in estimating the population mean μ
  - instead of the sample mean $\bar{X}$.
- By rearranging the terms of the above inequality, we find that with probability 0.95,

$$\bar{X} - 3 \leq \mu \leq \bar{X} + 3$$

  - This means that with probability 0.95, the population mean μ is in the interval $[\bar{X} - 3, \bar{X} + 3]$.
- The sample mean $\bar{X}$ is itself a random variable and changes from one sample to another.
  - Therefore, the above interval is not fixed.
  - With every new sample, we have a new value for $\bar{X}$, and as the result, we have a new interval.

Confidence Intervals for the Population Mean

- Theoretically, we could repeatedly sample *n* = 100 people, find the sample mean, and determine the interval.
- Then, the true population mean μ would fall within these intervals with probability 0.95.



- Suppose, for example, that we repeated this process 20 times to obtain 20 such intervals, as shown in the figure.
- In this figure, each sample mean is shown as a circle and the true (but unknown) population mean $\mu = 125$ as the dashed vertical line.
- Of 20 intervals, 19 (i.e., 95%) cover the true mean.

Confidence Intervals for the Population Mean

- In reality, however, we usually have only one sample of n observations, one sample mean $\bar{x}$, and one interval $[\bar{x} - 3,\ \bar{x} + 3]$ for the population mean μ.
- For the blood pressure example, suppose that we have a sample of $n = 100$ people and that the sample mean is $\bar{x} = 123$.
  - Therefore, we have one interval as follows:

$$[123 - 3,\ 123 + 3] = [120,\ 126].$$

- We refer to this interval as our 95% confidence interval for the population mean μ.

Confidence Intervals for the Population Mean

- In general, when the population variance σ^2 is known, the 95% confidence interval for the unknown population mean μ is obtained as follows:

  $$[\ \bar{x} - 2 \times \sigma/\sqrt{n}\ ,\ \bar{x} + 2 \times \sigma/\sqrt{n}\ ]$$

  where $\bar{x}$ is the specific value of the sample mean (i.e., observed sample mean) we obtain based on our sample.
- Alternatively, we say that the confidence level or confidence coefficient for the above interval is 0.95.

Confidence Intervals for the Population Mean

- Note that the above interval is only one of many possible intervals we could see.
- While we could assign a probability to all possible intervals based on $\bar{X}$ and say that 95% of them include the true value of the population mean, we cannot say the same thing for this specific interval based on $\bar{x}$.
- This specific interval is either one of those intervals that includes the true value of the population mean, or it is one of those intervals that do not.

Confidence Intervals for the Population Mean

- However, we are 95% confident that it belongs to the former set of intervals and includes the true value of the population mean.
- The 95% confidence refers to our degree of confidence in the *procedure* that generated this interval.
- If we could repeat this procedure many times, 95% of intervals it creates would include the true population mean.

Confidence Intervals for the Population Mean

- The multiplier 2 we used to obtain the above interval was derived from the 68–95–99.7 rule for normal distributions, which states that for a normally distributed random variable (in this case, $\bar{X}$), 95% of the observations fall within 2 SDs of the mean.
- If we want to increase our confidence level to 0.997, we use the multiplier 3 since 99.7% of observations fall within 3 SDs of the mean.
- Therefore, our 99.7% CI for the population mean is

$$[\ \bar{x} - 3 \times \sigma/\sqrt{n}\ ,\ \bar{x} + 3 \times \sigma/\sqrt{n}\ ]$$

- For the blood pressure example, the 99.7% CI is

$$[123 - 3\times1.5,\ 123 + 3\times1.5] = [118.5,\ 127.5]$$

- Alternatively, we say that the confidence level or confidence coefficient for the above interval is 0.997.

Confidence Intervals for the Population Mean

- For estimates at lower confidence level of 0.68, we use the multiplier 1 instead.
- Our 68% CI for the population mean is

$$[\ \bar{x} - \sigma/\sqrt{n}\ ,\ \bar{x} + \sigma/\sqrt{n}\ ]$$

- For the blood pressure example, the 68 % CI is

$$[123 - 1.5,\ 123 + 1.5] = [121.5,\ 124.5]$$

- Note that the length of this interval is smaller than the two previous interval estimates.

z-critical Values

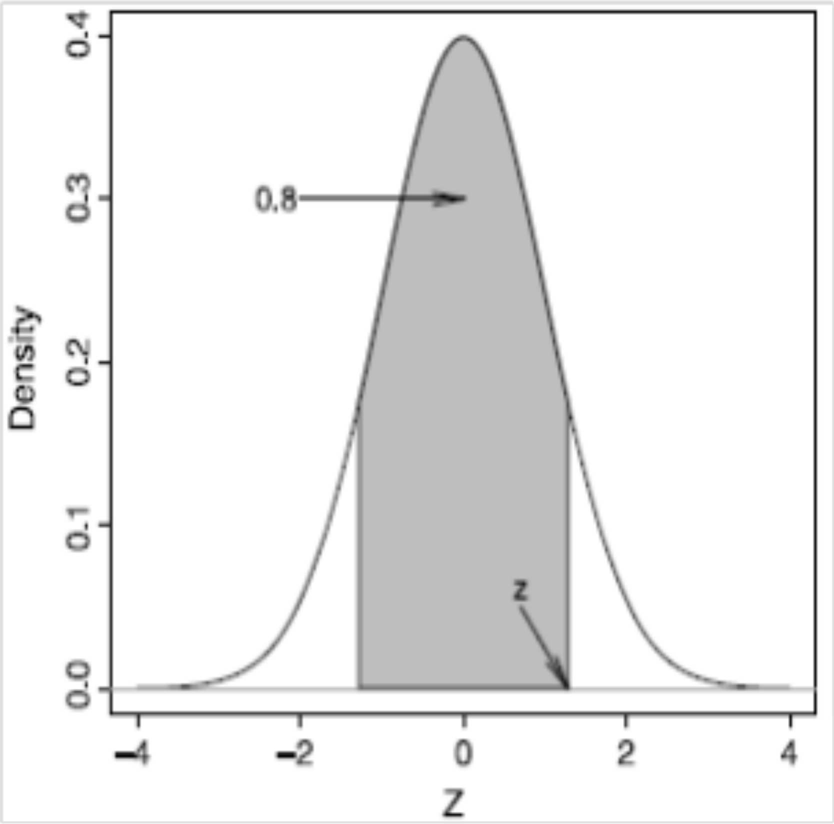
- In general, for a given confidence level, c, we use the standard normal distribution to find the value whose upper tail probability is $(1 - c)/2$.
- We refer to this value as the *z*-critical value for the confidence level of c.
- Then with the point estimate $\bar{x}$, the confidence interval for the population mean at c confidence level is

$$[\ \bar{x} - z_{\text{crit}} \times \sigma/\sqrt{n}\ ,\ \bar{x} + z_{\text{crit}} \times \sigma/\sqrt{n}\ ]$$

z-critical Values

- Suppose that we want to set the confidence level of our interval estimate for the population mean to 0.8.
- To find the corresponding multiplier, we need to find the number of units we need to move from 0 on each side so that the probability of the resulting interval becomes 0.8 based on the standard normal distribution.
- The next figure shows the probability density curve of $N(0, 1)$, which is known as the Z-curve.

z-critical Values



- The shaded area is 0.8,
  - which is the probability of the corresponding interval on the x axis.
- The upper end of this interval is shown as z.
  - Here, z is the number of units we need to move away from 0 so that the probability of the resulting interval is 0.8.
- That is, z is the multiplier needed to use to obtain 80% confidence intervals for population mean

z-critical Values

- Since the total area under the curve is 1, the unshaded area is $1 - 0.8 = 0.2$.
- Because of the symmetry of the curve around the mean, the two unshaded areas on the left and the right of the plot are equal,
  - which means that the unshaded area on the right-hand side is $0.2/2 = 0.1$.
- Therefore, the upper-tail probability of z is 0.1,
  - which is equal to $(1 - 0.8)/2$.

z-critical Values

- We can use R-Commander to find the value of z.
  - Click *Distributions* → *Continuous distributions* → *Normal distribution* → *Normal quantiles*.
  - Enter 0.1 for the Probabilities and select Upper tail.
- The result, shown in the Output window, is =1.28.
  - Therefore, we need to move $z = 1.28$ SDs from the mean on each side so that the probability of the resulting interval becomes 0.8.
- The 80% confidence interval for the population mean is

$$[\ \bar{x} - 1.28 \times \sigma/\sqrt{n}\ ,\ \bar{x} + 1.28 \times \sigma/\sqrt{n}\ ]$$

z-critical Values

- For the systolic blood pressure example, where $\bar{x} = 123$ and $\sigma/\sqrt{n} = 1.5$,
- we are 80% confident that the true mean blood pressure is in the interval

$$[123 - 1.28 \times 1.5,\ 123 + 1.28 \times 1.5] = [122.8,\ 123.2]$$

- We call the multiplier 1.28 the z-critical value, denoted as z_{crit}, for the 80% confidence interval.

z-critical Values

- We can follow similar steps to find the z-critical values for any other confidence level.
- For example, for 0.9 confidence level,

  $z_{\text{crit}} = 1.64$.
- For 0.95 confidence level, so far we have been using $z_{\text{crit}} = 2$.
- Following the above steps, you will find that a more accurate value is $z_{\text{crit}} = 1.96$, which is sometimes used instead of 2 to be more precise.

Standard error

- So far, we have assumed the population variance, σ^2, of the random variable is known.
  - This is an unrealistic assumption.
  - Almost always, we need to estimate σ^2 along with the population mean μ.
- For this, we use our sample of n observations to obtain the sample variance s^2 and sample standard deviation s.
  - As a result, the standard deviation for $\bar{X}$ is estimated to be $s/\sqrt{n}$
- We refer to $s/\sqrt{n}$ as the standard error of the sample mean $\bar{X}$ to distinguish it from $\sigma/\sqrt{n}$.
  - In general, we refer to the standard deviation of an estimator (e.g., $\bar{X}$) as its standard error (SE) if we have to use the data to estimate it.

Confidence Interval When the Population Variance is Unknown

- To find confidence intervals for the population mean when the population variance is unknown, we use
  - $SE = s/\sqrt{n}$ instead of $\sigma/\sqrt{n}$,
  - t_{crit} obtained from a t-distribution with $n-1$ degrees of freedom instead of z_{crit} based on the standard normal distribution.
- The confidence interval for the population mean at c confidence level is

$$[\, \bar{x} - t_{\text{crit}} \times s/\sqrt{n} \,,\, \bar{x} + t_{\text{crit}} \times s/\sqrt{n} \,]$$

Confidence Interval When the Population Variance is Unknown - Example

- Suppose that we have randomly selected seven newborn babies and recorded their heights (in inches) at the time of birth as follows:
  - Height: 18, 22, 19, 17, 20, 18, 15.
- The point estimates for μ and σ are $\bar{x} = 18.4$ and $s = 2.2$, respectively.
- The standard error (estimated SD) for the sample mean is $SE = 2.2/\sqrt{7} = 0.83$.

Confidence Interval When the Population Variance is Unknown - Example

- Suppose that we want to find the 90% confidence interval for the population mean, μ.
- Then, using the *t*-distribution with $7-1 = 6$ degrees of freedom, we need to find the *t*-critical value, t_{crit}, whose upper tail probability is $(1-0.9)/2 = 0.05$.
- In R-Commander,
  - click *Distributions* → *Continuous distributions* → *t distribution* → *t quantiles*.
  - Set the Probabilities to 0.05, the Degrees of Freedom to 6, and check Upper tail option.

Confidence Interval When the Population Variance is Unknown - Example

- The result, shown in Output window, is $t_{\text{crit}} = 1.94$, which is greater than $z_{\text{crit}} = 1.64$ based on the standard normal distribution.
- The 90% CI, therefore, is

$$\left[18.4 - 1.94 \times \frac{2.2}{\sqrt{7}}, 18.4 + 1.94 \times \frac{2.2}{\sqrt{7}}\right] = [16.8, 20.0]$$

- That is, at 0.9 confidence level, we estimate the mean of height for newborn babies to be between 16.8 and 20.0 inches.

Confidence Interval When the Population Variance is Unknown - Example

- In this example, if we knew $\sigma = 2.2$ instead of estimating it to be $s = 2.2$,
  - we would have used $z_{\text{crit}} = 1.64$ instead of $t_{\text{crit}} = 1.94$, and the interval would have been smaller.
- Everything else the same, using *t*-distribution instead of the standard normal leads to wider intervals.
  - This is the price we pay for the additional uncertainty due to the estimation of population variance (and SD) from the data.
- The *t*-distribution approaches the standard normal distribution as the sample size increases (i.e., the degree of freedom increase).
  - Therefore, the difference between the *z*-critical values and the *t*-critical values becomes negligible for very large sample sizes.

Using Central Limit Theorem for Confidence Interval

- So far, we have assumed that the random variable has normal distribution, so the sampling distribution of $\bar{X}$ is normal too.
- If the random variable is not normally distributed, the sampling distribution of $\bar{X}$ can be considered as approximately normal using (under certain conditions) the central limit theorem (CLT):
  - For large sample sizes, the CLT indicates that if the random variable X has the population mean μ and the population variance σ^2, then the sampling distribution of $\bar{X}$ is approximately normal with mean μ and variance σ^2/n
- Note that CLT is true regarding the underlying distribution of X so we can use it for random variables with Bernoulli and Binomial distributions too.

Confidence Intervals for the Population Proportion

- For binary random variables, we use the sample proportion to estimate the population proportion as well as the population variance.
- That is, the sample variance depends on the data through p and n only.
- Therefore, estimating the population variance does not introduce an additional source of uncertainty to our analysis,
  - so we do not need to use a *t*-distribution instead of the standard normal distribution.
- For the population proportion, the confidence interval is obtained as follows:

$$[p - z_{\text{crit}} \times SE, p + z_{\text{crit}} \times SE]$$

where $SE = \sqrt{p(1-p)/n}$

Confidence Intervals for the Population Proportion –Example-

- Suppose that we want to find the 95% CI for the population proportion of mothers who smoke during their pregnancy in the year 1986.
- Using the *birthwt* data set with $n = 189$, the estimate for this proportion is $\bar{x} = p = 0.39$.
- Using p, we estimate the population variance $p(1-p) = 0.39 \times 0.61 = 0.24$.
- The SE for the sample mean is

$$SE = \sqrt{p(1-p)/n} = \sqrt{(0.39 \times 0.61)/189} = 0.03$$

- The 95% CI is then ($z_{\text{crit}} = 1.96$, but round off to 2)

$$[p - z_{\text{crit}} \times SE, p + z_{\text{crit}} \times SE]$$

$$[0.39 - 2 \times 0.03, 0.39 + 2 \times 0.03] = [0.33, 0.45]$$

Confidence Intervals for the Population Proportion –Example-

- From the above confidence interval, we can find the confidence interval for the number of smoking pregnant women in the US during 1986.
- Supposing that there are currently $N = 4$ million pregnant women in the US,
- we find the 95% confidence interval for the number of smoking pregnant women as follows:

  $[0.33 \times 4000000,\ 0.45 \times 4000000] = [1320000,\ 1800000]$

Margin of Error

- For the above example, we can write the 95% CI for the population proportion of women who smoke during their pregnancy as follows:

$$0.39 \pm 2 \times 0.03.$$

- In this case, the term $2 \times SE = 2 \times 0.03 = 0.06$ is called the margin of error for 0.95 confidence level.
- In general, it is common to present interval estimates for c confidence level as

$$\text{Point estimate} \pm \text{Margin of error}$$

Margin of Error

- When the population variance σ^2 is known, the margin of error e is calculated as

$$e = z_{crit} \frac{\sigma}{\sqrt{n}}$$

- where z_{crit} is the multiplier obtained for the given confidence level c from the standard normal distribution.
- When the population variance is not known and we need to use the data to estimate it using the sample standard deviation, s, the margin of error is calculated as

$$e = t_{crit} \frac{s}{\sqrt{n}}$$

Sample Size Estimation

- Using the following equation for the margin of error:

$$e = z_{crit} \frac{\sigma}{\sqrt{n}}$$

- we can estimate the required sample size *n* for the assumed acceptable margin of error *e* as follows:

$$n = \left(\frac{z_{crit} \times \sigma}{e} \right)^2$$

Sample Size Estimation

- For example, let us find the appropriate sample size to estimate population mean for BMI.
- Suppose that we decide that the acceptable margin of error at confidence level 0.95 is 3.
- Further, suppose that, based on previous experience, we know that the BMI is roughly between 10 to 50.
- Therefore, we assume that σ is approximately

$$\sigma \approx \frac{range}{4} = \frac{max - min}{4} = \frac{50 - 10}{4} = 10$$

- Then the required sample size is:

$$n = \left(\frac{z_{crit} \times \sigma}{e}\right)^2 = \left(\frac{2 \times 10}{3}\right)^2 \approx 45$$

- Therefore, we need to measure the BMI of 45 people.