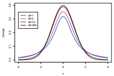
Chapter 14 Notes

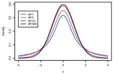
Confidence Intervals for One Mean
STP-231

Arizona State University



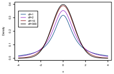
Statistical Inference

- We have essentially worked so far to use descriptive measures to discuss characteristics of observed data. We have discussed theoretical parameters of a population
- We have also discussed the limitations of using samples to estimate populations. Now, we more thoroughly quantify this
- **Statistical Inference** describes methods for making predictions about a population based on information collected from a sample
- For example, we will look at statistical estimation, confidence intervals, and hypothesis testing



Statistical Estimation

- In statistical estimation, we make inferences about data to:
- Determine an estimate of some value of the population (i.e. a statistic to estimate a parameter)
- Determine how precise the estimate is (i.e. figure out the standard deviation of our estimator/statistic)
- A **point estimate** is a value of a statistic that is used as an estimate of a parameter, i.e. our “best guess”

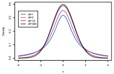


Example 1

As part of a larger study of body composition, researchers captuered 14 Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in cm^2)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wing size	33.9	33.0	30.6	36.6	36.5	34.0	36.1	32.0	28.0	32.0	32.2	32.2	32.3	30.0

Calculate the mean and standard deviation



Example 1

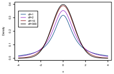
As part of a larger study of body composition, researchers captuered 14 Monarch butterflies at Oceano Dunes State Park in California and measured wing area (in cm^2)

Index	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Wing size	33.9	33.0	30.6	36.6	36.5	34.0	36.1	32.0	28.0	32.0	32.2	32.2	32.3	30.0

Calculate the mean and standard deviation Let Y be the random variable representing the wing area. What are the units on each?

$$\overline{Y} \approx 32.81 \qquad s \approx 2.48$$

both are in units cm^2

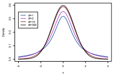


Standard Error of the Mean

- The standard deviation of $\overline{Y}$ is $\frac{\sigma}{\sqrt{n}}$ Why? Because $\text{Var}(a \cdot X) = a^2\text{Var}(X)$, that is:

$$\text{Var}(\overline{Y}) = \text{Var}(\sum_{i=1}^{n} \frac{y_i}{n}) \text{ (Since } y_i \text{ identically distributed)} \quad (1)$$

$$= \frac{n}{n^2}\text{Var}(Y) \implies \sigma_{\overline{Y}} = \frac{\sigma}{\sqrt{n}} \quad (2)$$



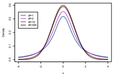
Standard Error of the Mean

- Say we are using $\overline{Y}$ as a point estimate of μ. We use S as an estimate of σ

- Similarly, $\frac{\sigma}{\sqrt{n}}$, the standard deviation of the sampling distribution can be estimated with $\frac{s}{\sqrt{n}}$ We call this the

  standard error of the mean,

$$\mathrm{SE}_{\overline{Y}} = \mathrm{SE} = \frac{s}{\sqrt{n}}$$



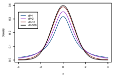
Standard Deviation vs Standard Error

- Standard deviation: Inherent variation (or within sample variation), refers to spread of the sample values. What we "expect" within a sample

- Standard error: Uncertainty due to sampling error in the mean of the data

- Measuring reliability that our calculated sample means are close to the actual population mean. Factors in within sample variation and the sample size (sample to sample variation)

Confidence Intervals

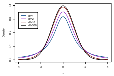


Confidence Intervals

- **Confidence Interval**: Also known as CI, is an interval of values obtained from a point estimate of a parameter
- Usually takes the form

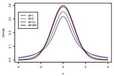
$$\text{point estimate} \; \pm \; \text{margin of error}$$

- **Confidence Level**: How sure we are the parameter lies in the confidence interval
- **Margin of error**: Some measure of the sampling error
- **Confidence – interval estimate**: The confidence level and interval



Confidence Level

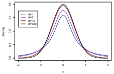
- **Confidence level**: the confidence we have that the parameter lies in the confidence interval
- Written in the form $1 - \alpha$, where α is a number between 0 and 1
- α is called the significance level
- $1 - \alpha$ is the success rate of the method that produces the interval



Example 2

The number of offspring of Eastern Cotton Mouth snakes is believed to be smaller b/c of human encroachment. 44 female snakes were randomly samples and the number of offspring from each snake was counted. The data is below:

5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6



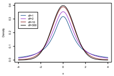
Example 2

The number of offspring of Eastern Cotton Mouth snakes is believed to be smaller b/c of human encroachment. 44 female snakes were randomly samples and the number of offspring from each snake was counted. The data is below:

5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6

This yields $\overline{Y} = 7.59$ young/litter and $\sigma = 2.4$ (assuming we know standard deviation)

- We know $\overline{Y}$ is a point estimate for μ. How likely is it that they are exactly equal? We can create a range of values where we have some confidence that μ will fall into

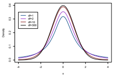


Example 2

The number of offspring of Eastern Cotton Mouth snakes is believed to be smaller b/c of human encroachment. 44 female snakes were randomly samples and the number of offspring from each snake was counted. The data is below:

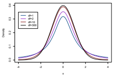
5	12	7	7	6	8	12	9	7	4	9
6	12	7	5	6	10	3	10	8	8	12
5	6	10	11	3	8	4	5	7	6	11
7	6	8	8	14	8	7	11	7	5	6

- Consider two standard deviations away from $\overline{Y}$. We are 95.44% confident that the mean falls within these limits
- $\overline{Y} \pm 2 \times \frac{\sigma}{\sqrt{n}} = 7.59 \pm 2 \times \frac{2.4}{\sqrt{44}}$, which is $(6.87, 8.31)$
- **Interpretation** “If we were to do this study 100 times, approximately 4-5 of them would not contain the true population mean



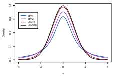
Interpreting a Confidence Interval

- The confidence level is the success rate of the method that produced the interval. We don't know if the $1-\alpha$ CI for particular sample will be one of these successes. We do know:
- $(1-\alpha)\times 100\%$ if intervals will be success
- $\alpha \times 100\%$ will not be
- Example: To say we are 95% confident that the unknown value of μ falls within a 1-0.05 confidence interval is saying: "We got these intervals using a method that gives correct results 95% of the time."
- **The confidence level should never be interpreted as the probability that a parameter is within a specific confidence interval"
- cool link



Example 3: Incorrect Interpretation

- We obtain a 90% confidence interval for a population parameter $\mu : (8.13, 10.4)$
- Based on one sample alone. We are not interested in quantifying $\Pr(8.13 < \mu < 10.4)$ because μ is fixed and so is the sample for this confidence interval in particular
- The probability mentioned above is either 0 or 1, the mean is in the interval or it is not
- We do consider how often this method does contain the population parameter
- In this way the confidence level is the proportion of successes that we would count over many repetitions of this method
- “90% of the time the mean would be contained in an interval formed this way

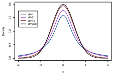


Confusing

- Confidence intervals confuse a lot of people. Case in point:

> The uncertainty around a point estimate can be small or large. Scientists represent this uncertainty by calculating a range of possibilities, which they call a confidence interval. One way of thinking of a confidence interval is that we can be 95 percent confident that the efficacy falls somewhere inside it. If scientists came up with confidence intervals for 100 different samples using this method, the efficacy would fall inside the confidence intervals in 95 of them.

- These are conflicting interpretations (from the New York Times)



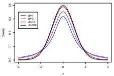
Why is Bayesian Statistics Different?

Assume we take a test that is 97% true positive and 2% false positive. Assume 60% of people taking the test have the disease of interest. What is the probability that if someone tests positive twice in a row they actually have the disease?
Let P be the random variable having the positive test. Then from bayes rule:

$$\begin{aligned}\Pr(D \mid P) &= \frac{\Pr(P \mid D)\Pr(D)}{\Pr(P \mid D) + \Pr(N \mid D)\Pr(N)} \\ &= \frac{0.97 \times 0.60}{0.97 \times 0.60 + 0.02 \times 0.4} = 0.9864\end{aligned}$$

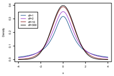
Now, for the second text, $\Pr(D)$ becomes 0.9864, i.e. we are updating our *prior* belief, yielding

$$\Pr(D \mid P, P) = \frac{0.97}{0.97 \times 0.9864 + 0.02 + 0.01364} = 0.9997$$



Bayesian interpretation part 2

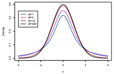
In terms of Bayesian statistics, the parameter is not **fixed** but itself a random variable which is updated based on what our prior belief is (probability of D we assumed before seeing any data)



Confidence Interval for the Pop. Mean (σ known)

- Assume a simple random sample
- Also assume the population is normally distributed and we have a large sample
- If we know the population standard deviation, then the $(1-\alpha)\times 100\%$ confidence interval for the population is:

$$\overline{Y} \pm z_{\alpha/2} \times \frac{\sigma}{\sqrt{n}}$$



Intuition

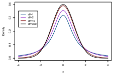
- Let C be any confidence level. We call $z^* = z_{\alpha/2}$ and $-z^* = -z_{\alpha/2}$ the critical values. They are the values that provide probability C within their range under the curve.
- If we start at the sample mean and move outward by z^* standard deviations on either side, we get an interval that contains the population mean μ in a proportion C of all samples
- Recall:

$$z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \implies \mu = \overline{Y} - z \times \frac{\sigma}{\sqrt{n}}$$

solve:

$$z_{\alpha/2} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \text{ and } -z_{\alpha/2} = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}}$$

for μ



Visual Intuition

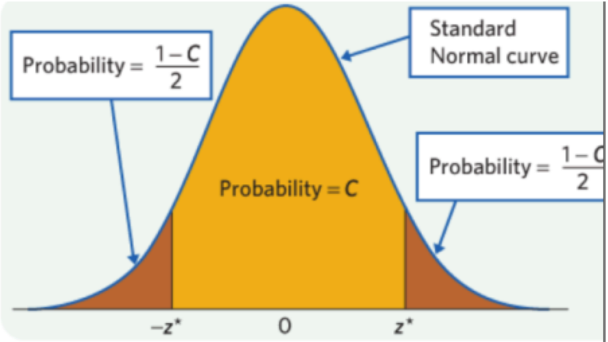
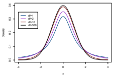


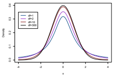
Figure 14.3
Baldi/Moore, *The Practice of Statistics in the Life Sciences*, 4e, © 20
W. H. Freeman and Company



Example 4

Researchers are trying to predict the mean body temperature of humans. They obtained the body temperature of 93 humans. The mean body temperature if 98 degrees fahrenheit. Assume population standard deviation is 0.63 degrees fahrenheit (maybe we know from past studies). Find a 95% confidence interval for the mean body temperature. What is the correct interpretation?

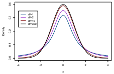
When σ is unknown



When σ is unknown

- This is more realistic
- In this case, we use the sample standard deviation as an estimate of the truth, i.e. S instead of σ
- Instead of the standard z, we now use the studentized version of $\overline{Y}$, which means we use a t-distribution to approximate a normal

$$z = \frac{\overline{Y} - \mu}{\sigma/\sqrt{n}} \longrightarrow t = \frac{\overline{Y} - \mu}{S/\sqrt{n}}$$

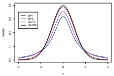


Student's t-distribution

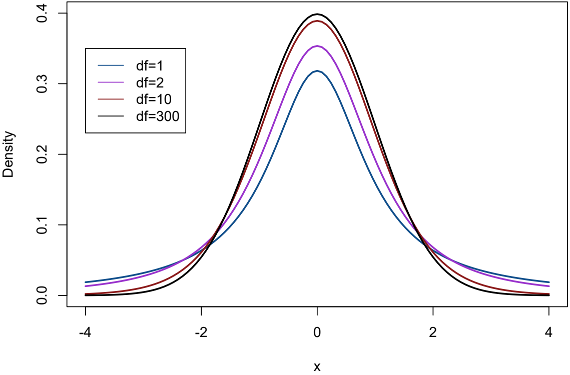
- Bell-shaped curve centered at 0, with thicker tails
- Uses degrees of freedom, where df=$n-1$ for n data points in the sample
- As n increases, looks more and more like normal distribution
- In all it's glory, the distribution is given by (where $\nu = n-1$ degrees of freedom) Wiki article

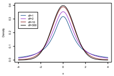
$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)\left(1+\frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}}$$

$\Gamma(n) = (n-1)!$ if integer. O.w. $\Gamma(x) = \frac{1}{x}\Gamma(x+1)$



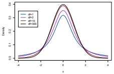
Visual t-distribution





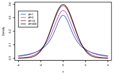
How to make the plot

```
#make the curves
curve(dt(x, df=1), lwd=2,lty=1,from=-4, to=4, col='dodgerblue4', ylim=c(0, 0.4), ylab='Density')
#need add=T to get multiple.
curve(dt(x, df=2), lwd=2,lty=1,from=-4, to=4, col='darkorchid', add=T, ylim=c(0, 0.4), ylab='Density')
curve(dt(x, df=10), lwd=2, lty=1,from=-4, to=4, col='firebrick4', add=TRUE, ylim=c(0,0.4), ylab='Density')
curve(dt(x, df=300), lwd=2,lty=1,from=-4, to=4, col='black', add=TRUE, ylim=c(0, 0.4), ylab='Density')
#add legend
legend(-4, .35, legend=c("df=1", "df=2",  "df=10", "df=300"),
col=c("dodgerblue4","darkorchid",  "firebrick4", "black"), lty=c(1,1,1,1), cex=1.)
```



More on the t-distribution

- Use the t-table for the area under the curve
- Area under the curve is the right tail for df=$n-1$ where $1-\alpha$ is the confidence level and α is the significance level. Find:
- $t_{0.10}$ for $n=10$
- $t_{0.05}$ for $n=26$
- $t_{?}=1.948$ for $n=11$
- $t_{?}=1.973$ for $n=10$

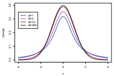


CI for pop mean (σ unknown)

We make the following assumptions:

- Take a simple random sample (SRS)
- Normal population or a large sample
- σ is unknown
- For a confidence level of $1 - \alpha$, we can generate a confidence interval for μ by

$$\overline{Y} \pm t_{\alpha/2} \times \frac{S}{\sqrt{n}}$$

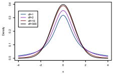


Example 4

Snakes deposit chemical trails as they travel through their habitats. These trails are often detected and recognized by lizards which are potential prey. The ability to recognize their predators via tongue flicks can often mean life or death for lizards. Scientists were interested in quantifying the responses of the common lizard to natural predator cues to determine whether the behavior is learned or congenital. Seventeen juvenile common lizards were exposed to the chemical cues of the viper snake. Their responses in number of tongue flicks per twenty minutes are:

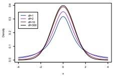
425	510	629	236	654	200
276	501	811	332	424	674
676	694	710	662	663	

Construct a 90% confidence interval and interpret



Example 5

The subterranean coruro (Spalacopus cyanus) is a social rodent that lives in large colonies in underground burrows that can reach lengths of up to 600 meters. A sample of 51 burrows had an average depth of 15.05 centimeters with a sample standard deviation of 2.50 centimeters. Construct a 95 % confidence interval for the mean burrow depth for all subterranean coruros.

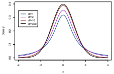


Example 5

The subterranean coruro (Spalacopus cyanus) is a social rodent that lives in large colonies in underground burrows that can reach lengths of up to 600 meters. A sample of 51 burrows had an average depth of 15.05 centimeters with a sample standard deviation of 2.50 centimeters. Construct a 95 % confidence interval for the mean burrow depth for all subterranean coruros.

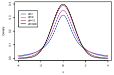
$$15.05 \pm \frac{2.50}{\sqrt{51}} \times t_{0.025,51-1}$$
$$= 15.05 \pm -2.01 \times 0.35 \implies 14.35 < \overline{Y} < 15.75$$

Planning Study Sample Size



Planning a Study for Estimating μ

- Find sufficient sample size to estimate the parameter with an acceptable confidence level
- After spending time and resources it is upsetting to realize there is an insufficient amount of data to draw a reasonable conclusion
- Reduce variability: Example, for a Breast Cancer study on five-year survival rates the data on the patients could be organized into groups such as the following:
- Stage I-IV diagnosis
- Pre-menopausal or post-menopausal
- Estrogen, progesterone, or HER2 negative/positive
- Chemotherapy-yes or no
- Radiation-yes or no



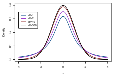
Determining Sample Size

- What sample size will be sufficient to achieve a desired degree of precision in estimation of the population mean?
- Use the standard error as our measure of precision

$$\text{SE}_{\overline{Y}} = \frac{S}{\sqrt{n}}$$

- The required sample size is then determined from the following equation

$$\text{Desired Standard Error} = \frac{\text{Guess standard deviation}}{\sqrt{n}}$$

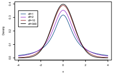


Example 6

Assume we have the following data on butterfly wings.

$$\overline{Y} = 32.81 \text{ cm}^2$$
$$S = 2.48 \text{ cm}^2$$
$$\text{SE} = 0.66 \text{ cm}^2$$

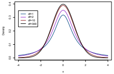
Suppose the researcher is now planning a new study of butterflies and has decided that it would be desirable that the SE be no more than 0.4 cm^2. What n would we need?



Margin of Error, Precision, Sample Size

- Length of a confidence interval is a measure of the precision with which $\overline{Y}$ estimates μ
- For a fixed confidence level, increasing the sample size improves the precision
- If a confidence level and margin of error is given, then the appropriate sample size needed to meet those specifications must be determined from the formula:

$$E = t_{\alpha/2} \times \frac{S}{\sqrt{n}}$$



Example 7

Physical therapy students during their graduate-school years were studied by the College of Health at the University of Nevada, Las Vegas. The researchers were interested in the fact that, although graduate physical therapy students are taught the principles of fitness, some have difficulty finding the time to implement those principles. Assuming that percent body fat of female graduate physical-therapy students is normally distributed with standard deviation 4.10 percent body fat. Determine the sample size required to have a margin of error of 1.25 percent body fat with 95% confidence level.