Statistical Intervals Based on a Single Sample (7.1 - 7.3)

1. Introduction: Basic Definitions

   (a) **Inerval Estimate:** When the value of a population parameter is estimated, an alternative to reporting a single value is to report an entire interval of plausible values for the population parameter. This interval, called a **confidence interval**, has a high probability of containing the true value of the population parameter being estimated.

2. Basic Properties of Confidence Intervals

   (a) Consider this simple problem situation:

   - Population Distribution of X is known to be: $N(\mu, \sigma^2)$
   - μ is unknown, we want to estimate it's value.
   - σ's value is known.
   - Sample observations $X_1, X_2, ..., X_n$ are the result of a random sample from the above population.

   (b) Development of a $100(1-\alpha)\%$ Confidence Interval for the population mean, μ, begins with the sampling distribution of $\bar{X}$. $\bar{X}$ is the sample statistic that is an estimator of μ. Confidence Intervals are constructed to contain the population mean, μ, with high probability.

Population	Sample
$X \sim N(\mu, \sigma^2)$	$\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
	$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

   i. The basic probability statement for the confidence interval is

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

   ii. Substituting $(\frac{\bar{x}-\mu}{\sigma/\sqrt{n}})$ for Z :

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq z_{\frac{\alpha}{2}}) = 1 - \alpha$$

   iii. Algebraic rearrangements to isolate μ in center

$$P(-z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

   iv. When the value of σ is known, given values for $\bar{x}$ and n, the $100(1-\alpha)\%$ confidence interval is

$$(\bar{x} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \; , \; \bar{x} + z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$$

(c) **Confidence Level:** The confidence level of a confidence interval is a measure of the degree of reliability of the interval.

- **$100(1-\alpha)$% Confidence Interval for μ:** After drawing random sample $X_1, X_2, ..., X_n$, first compute the sample mean $\bar{x}$ as a point estimate of μ, the population mean. Then, a confidence interval for μ can be expressed by its lower and upper bound in parentheses.

  $$(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \; , \; \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}})$$

  A confidence level of 95%, means that $(1-\alpha) = .95$ and that $\alpha/2 = .025$. A confidence level of 90%, means that $(1-\alpha) = .90$ and that $\alpha/2 = .050$.

- Interpreting Confidence Intervals: If we sample the population many many times, in the long run, $100(1-\alpha)$% of our computed confidence intervals (CI's) will contain μ, the other 100α% will not.
- A smaller CI width indicates a more precise estimate of μ. The interval half-width is sometimes called the bound on the error of estimation.
- The CI is centered on the sample mean, $\bar{x}$. It's width depends on both α and n.
- As α increases $Z_{\alpha/2}$ gets smaller and the interval width decreases.
- As n increases $\frac{\sigma}{\sqrt{n}}$ gets smaller and the interval width decreases.
- Commonly used values for $Z_{\alpha/2}$ are tabulated below.

Confidence Level	$(1-\alpha)$	α	$\alpha/2$	$Z_{\alpha/2}$
99%	.99	.01	.005	2.575
95%	.95	.05	.025	1.960
90%	.90	.10	.050	1.645
80%	.80	.20	.100	1.280

3. Example: Suppose Jane weighs herself once a week for 12 weeks and records the following weights in pounds.

```
145.1 152.3 143.2 147.8 149.4 147.2
151.7 146.3 149.3 150.2 151.2 152.7
```

If her weight follows a normal distribution with standard deviation, $\sigma = 3$, compute a 90% confidence interval for her mean weight.

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

- $\bar{x} = \frac{\sum x_i}{n} = 148.87$ and $n = 12$
- $(1-\alpha) = .90$, so that $\alpha/2 = .050$
  and $z_{\alpha/2} = z_{.050} = 1.645$ from the table.

- $U = 148.8 + 1.645(\frac{3}{\sqrt{12}}) = 148.8 + 1.425 = 150.2$
- $L = 148.8 - 1.645(\frac{3}{\sqrt{12}}) = 148.8 - 1.425 = 147.4$
- 90% Confidence Interval for μ: $(L, U) = (147.4, 150.2)$

4. Precision and Sample Size

   - Sample Size Determination (σ known): The width of the confidence interval developed above depends on sample size, n. As sample size increases, the confidence interval width decreases.
   - Confidence interval width: $w = 2(z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}})$
   - The sample size necessary to insure an interval width, w is:

$$n = (\frac{2z_{\alpha/2}\sigma}{w})^2$$

   - When computing n, always round up to the next whole number.
   - The smaller the desired w, the larger n has to be.

5. Large-Sample Confidence Intervals for Population Mean and Proportion

   (a) When value of σ is unknown

   - When the sample size is large, by invocation of the Central Limit Theorem, the sampling distribution of the sample mean, $\bar{X}$, is at least approximately normally distributed even when the population distribution is not normal.
   - The sampling distribution for $\bar{X}$, estimator for μ, is again the starting point for developing a confidence interval for μ. When the value of σ is not known, its value must be estimated using the sample standard deviation, s. Using s instead of σ creates a few complications in the use of the previously derived sampling distribution for $\bar{X}$.

Std. Variable	Distribution	Condition
$Z = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$	always $\sim N(0,1)$	σ known
$Z = \frac{\bar{x}-\mu}{s/\sqrt{n}}$	approx. $\sim N(0,1)$ for large n, $(n > 40)$	σ unknown large sample
$t = \frac{\bar{x}-\mu}{s/\sqrt{n}}$	$\sim t(\nu)$ where $\nu = n - 1$ for small n, $(n < 40)$	σ unknown small sample

   - As a consequence of the above, the confidence interval for μ is computed differently for large samples, $(n > 40)$, and small samples, $(n < 40)$, when the value of σ is unknown.

(b) Large-Sample Interval for μ

- If n is sufficiently large, $(n > 40)$, standard rv Z has approximately a standard normal distribution, $N(0,1)$, when σ is replaced by s.

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

- The $100(1-\alpha)\%$ Confidence Interval for μ is then

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- This is a large-sample confidence interval for μ and is valid regardless of the shape of the population distribution. $(n > 40)$ is generally sufficient justification for use of this interval.

(c) Example: Suppose Jane continues weighing herself once a week for an entire year and records the weights in pounds each week. The following are summary statistics from her past year.

- Sample Size: $n = 52$
- Sample Mean: $\bar{x} = \frac{\sum x_i}{n} = 148.87$
- Sample Standard Deviation: $s = 3.00$

If her weight follows a normal distribution with unknown standard deviation, compute a 99% confidence interval for her mean weight.

$$\bar{x} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$$

- $(1-\alpha) = .99$, so that $\alpha/2 = .005$
  and $z_{\alpha/2} = z_{.005} = 2.575$ from the table.

- $U = 148.87 + 2.575(\frac{3}{\sqrt{52}}) = 148.87 + 1.07 = 149.94$

- $L = 148.87 - 2.575(\frac{3}{\sqrt{52}}) = 148.87 - 1.07 = 147.80$

- 99% Confidence Interval for μ: $(L, U) = (147.8, 149.9)$

(d) General Large-Sample Confidence Interval

i. When $\hat{\theta}$ is an estimator of population parameter θ, if $\hat{\theta}$:
   - has approximately a normal distribution
   - is approximately an unbiased estimator of θ
   - has an available expression for $\sigma_{\hat{\theta}}$, the standard deviation of $\hat{\theta}$

ii. Then a confidence interval for θ takes the following general form:

$$\hat{\theta} \pm Z_{\alpha/2} \cdot \sigma_{\hat{\theta}}$$

iii. This is the general form for a large-sample confidence interval for θ which applies to more than just μ.

6. Confidence Interval for Population Proportion

(a) Given a large population of size N, containing a count of X_p successes, the population proportion of successes, p, is:

$$p = \frac{X_p}{N}$$

(b) To estimate p, when X_p and p are unknown, a random sample of size n is taken (without replacement) from the population and rv X is the count of successes observed in the sample.

(c) The sample proportion, $\hat{p}$, is determined from the sample as $\frac{X}{n}$. It is our MLE estimator of p.

- When n is small compared to N, X can be regarded as a binomial rv with

$$E(X) = np$$

$$Var(X) = np(1-p)$$

- If n is large, so that $np \geq 10$ and $n(1-p) \geq 10$, then X is at least approximately normally distributed:

$$X \sim N(np,\ np(1-p))$$

- Since $\hat{p} = \frac{1}{n}X$:

$$E(\hat{p}) = \frac{1}{n}E(X) = p$$

$$Var(\hat{p}) = \left(\frac{1}{n}\right)^2 Var(X) = \frac{p(1-p)}{n}$$

- So that

$$\hat{p} \sim N(p,\ p(1-p)/n)$$

$$Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

- Basic probability statement for confidence interval for p is then:

$$P(-z_{\frac{\alpha}{2}} \leq Z \leq z_{\frac{\alpha}{2}}) \approx 1-\alpha$$

- Substituting $(\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}})$ for Z :

$$P(-z_{\frac{\alpha}{2}} \leq \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\frac{\alpha}{2}}) \approx 1-\alpha$$

- Rearrangements to isolate p in center result in a quadratic equation in p.
- This equation has been solved to provide the following upper (U) and lower (L) confidence interval bounds for a $100(1-\alpha)\%$ CI for population proportion, p. Here $\hat{q} = 1-\hat{p}$.

$$U = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}$$

and

$$L = \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}$$

- To guarantee a specified interval width, w, we must choose sample size n from the relationship shown below with the largest value:

$$n = \frac{2Z_{\alpha/2}^2 \hat{p}\hat{q} - Z_{\alpha/2}^2 w^2 \pm \sqrt{4Z_{\alpha/2}^4 \hat{p}\hat{q}(\hat{p}\hat{q} - w^2) + w^2 Z_{\alpha/2}^4}}{w^2}$$

- For a specified w, but unknown $\hat{p}$, to be conservative use $\hat{p} = 0.5$ as it produces the largest value for $\hat{p}\hat{q} = 0.25$

- When the value of n is quite large, the above CI for p can be simplified due to the following:

$$\hat{p} >> \frac{z_{\alpha/2}^2}{2n} \quad ; \quad \frac{\hat{p}\hat{q}}{n} >> \frac{z_{\alpha/2}^2}{4n^2} \quad ; \quad 1 >> z_{\alpha/2}^2/n$$

- The approximate CI is then:

$$\hat{p} \pm Z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

- Note that this is the general form of the large sample confidence interval presented earlier.
- Using this approximation, the sample size, n, needed to guarantee a specified interval width, w, can now be computed as:

$$n \approx \frac{4Z_{\alpha/2}^2 \hat{p}\hat{q}}{w^2}$$

- As above, use $\hat{p} = 0.5$ to determine n in a conservative fashion.

(d) Example: When a random sample of 37 suspension football helmuts were subjected to a specific impact test, 24 of them showed damage. Let p represent the proportion of all helmets of this type that would show damage when subjected to the above impact test.

i. Calculate a 99% confidence interval for p.
  - $\hat{p} = \frac{24}{37} = 0.6486$
  - The 99% CI for p is

$$U = \frac{0.6486 + \frac{(2.58)^2}{2(37)} + 2.58\sqrt{\frac{(0.6486)(0.3514)}{37} + \frac{(2.58)^2}{4(37)^2}}}{1 + \frac{(2.58)^2}{37}} = \frac{0.7386 + 0.2216}{1.1799} = 0.814$$

$$L = \frac{0.6486 + \frac{(2.58)^2}{2(37)} - 2.58\sqrt{\frac{(0.6486)(0.3514)}{37} + \frac{(2.58)^2}{4(37)^2}}}{1 + \frac{(2.58)^2}{37}} = \frac{0.7386 - 0.2216}{1.1799} = 0.438$$

  - The CI is: (0.438 , 0.814)

ii. What sample size, n, would be required for a 99% CI width to be at most 0.10?

$$n = \frac{2(2.58)^2(0.25) - (2.58)^2(0.01) \pm \sqrt{4(2.58)^4(0.25)(0.25 - 0.01) + 0.01(2.58)^4}}{0.01}$$
$$= \frac{3.261636 \pm 3.3282}{0.01} \approx 659$$

7. Intervals Based on Normal Population Distribution

(a) When the population of interest is normal, then $X_1, X_2, ..., X_n$ constitutes a random sample from a normal distribution with unknown μ and σ. If the sample size is small, $n < 40$, and s is used to estimate σ, then the sampling distribution for $\bar{X}$, when standardized, becomes a T statistic (rv) which follows a t-distribution with $\nu = n - 1$ degrees of freedom.

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(\nu)$$

(b) Properties of t Distributions

- Each $t(\nu)$ curve is bell-shaped and centered at 0, like the standard normal distribution, $N(0, 1)$.
- $t(\nu)$ curves tend to be a bit shorter and fatter than the standard normal distribution.
- As ν, the number of degrees of freedom, increases the spread of the $t(\nu)$ decreases.
- As $\nu \to \infty$, the $t(\nu)$ curve becomes identical to the standard normal curve, $N(0, 1)$. The z curve is a $t(\nu)$ curve with $\nu = \infty$.
- Critical Values of t: When the area under the $t(\nu)$ curve to the right of some T value, say t_{crit}, is equal to α, then $t_{crit} \equiv t_{\alpha,\nu}$ is called a critical value of t.

(c) Small Sample Confidence Interval for μ (σ unknown)

- If n is small, $(n < 40)$, the standardized variable Z does not follow a standard normal distribution, $N(0, 1)$, when σ is replaced by s. Rather, it follows a t-distribution with $\nu = n - 1$ degrees of freedom and is designated as t instead of Z.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t(\nu)$$

- The $100(1 - \alpha)\%$ Confidence Interval for μ is

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

- This is the small sample confidence interval for population mean, μ.

(d) Example: Suppose we evaluate vitamin C levels (mg/100 gm) in 8 batches of corn soy blend(CSB) from a production run and get:

```
26 31 23 22 11 22 14 31
```

Find a 95% confidence interval for the mean vitamin C content of CSB produced during this run.

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$

- $\bar{x} = \frac{\sum x_i}{n} = 22.50$, $s = 7.19$ and $n = 8$

- $(1 - \alpha) = .95$, $\alpha/2 = .025$, $\nu = n - 1 = 7$ and $t_{\alpha/2,\nu} = t_{.025,7} = 2.365$ from the table.

- $U = 22.50 + 2.365(\frac{7.19}{\sqrt{8}}) = 22.50 + 6.012 = 28.5$
- $L = 22.50 - 2.365(\frac{7.19}{\sqrt{8}}) = 22.50 - 6.012 = 16.5$
- 95% Confidence Interval for μ: $(L, U) = (16.5, 28.5)$

(e) Prediction Interval (PI) for a Single Future Observation, X_{n+1}, to be selected from a normal population distribution is

$$\bar{x} \pm t_{\alpha/2,n-1} \cdot s\sqrt{1 + \frac{1}{n}}$$

here the prediction level is $100(1 - \alpha)\%$.

(f) Example: Compute the 95% PI for X_{n+1} in the above example where $\bar{x} = 22.50$, $s = 7.19$, $t_{.025,7} = 2.365$, and $n = 8$.

- $U = 22.50 + 2.365(7.19)\sqrt{1 + \frac{1}{8}} = 22.50 + 17.004(1.061) = 40.54$
- $L = 22.50 - 2.365(7.19)(1.061) = 22.50 - 18.042 = 4.46$
- 95% Prediction Interval for X_{n+1}: $(L, U) = (4.5, 40.5)$
- As one might expect, this PI for X_{n+1} is quite a bit wider than the CI computed above for the mean, μ.