1036: Probability & Statistics

Lecture 9 - One- and Two-Sample Estimation Problems
Statistical Inference

• Estimation
  - to estimate the population parameters
  - Classical
    • Based on random sample
  - Bayesian
    • Based on prior subjective knowledge about the prob. distribution and random sample

• Tests of hypothesis
  - an assertion or conjecture concerning one or two populations
Unbiased Estimator

• An estimator may not expect to estimate the exact value of population parameter
  - But hope that it is not fall off...

• Unbiased estimator
  - A statistic is said to be unbiased if its sampling distribution has a mean equal to the parameter estimated

\[ \mu_\theta = E(\hat{\Theta}) = \theta \]
Show that $S^2$ is an unbiased estimator of the parameter $\sigma^2$

\[
E(S^2) = E \left( \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1} \right) = \frac{1}{n-1} E \left( \sum_{i=1}^{n} [(X_i - \mu) - (\bar{X} - \mu)]^2 \right)
\]

\[
= \frac{1}{n-1} E \left[ \left( \sum_{i=1}^{n} (X_i - \mu)^2 \right) - n(\bar{X} - \mu)^2 \right] = \frac{1}{n-1} \left[ \sum_{i=1}^{n} E(X_i - \mu)^2 - nE(\bar{X} - \mu)^2 \right]
\]

\[
= \frac{1}{n-1} \left( \sum_{i=1}^{n} \sigma_{X_i}^2 - n\sigma_{\bar{X}}^2 \right)
\]

However,

\[
\sigma_{X_i}^2 = \sigma^2 \quad \text{for } i = 1, 2, \ldots, n \quad \text{and} \quad \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n}
\]

Therefore,

\[
E(S^2) = \frac{1}{n-1} \left( \sum_{i=1}^{n} \sigma^2 - n \frac{\sigma^2}{n} \right) = \sigma^2
\]
Most Efficient Estimator

- The one with the smallest variance among all possible unbiased estimators of some population $\theta$ is called the most efficient estimator of $\theta$.

\[ \begin{align*}
\hat{\Theta}_1 & \quad \text{unbiased} \\
\hat{\Theta}_2 \\
\hat{\Theta}_3
\end{align*} \]

Figure 9.1 Sampling distributions of different estimators of $\theta$. 
Interval Estimate

- Even most efficient unbiased estimator not likely to estimate exactly correctly.
- Although accuracy increases with large samples, no reason why a point estimate from a sample should exactly equal the population parameter.
- One way to handle this error is through an interval estimate
  - Example: sample mean = 540
    - Confidence interval: $520 < m < 560$
    - Since $s_x^2 = \sigma^2/n$, accuracy should increase with $n$ and interval size should decrease.
    - Range of interval indicates accuracy of the point estimate
Interpretation of Interval Estimates

- If the confidence interval is \( \hat{\theta}_L < \theta < \hat{\theta}_U \)
- The probability that the mean is within the range can be stated as:

\[
P(\hat{\theta}_L < \theta < \hat{\theta}_U) = 1 - \alpha
\]

- We would state that there is a \((1-\alpha)\times100\%\) confidence interval of \( \hat{\theta}_L < \theta < \hat{\theta}_U \)

- Ideally, predict narrow range with high degree of confidence
Estimating the Mean

- The sampling distribution of $\overline{X}$ is centered at $\mu$ and $\sigma^2_{\overline{X}} = \sigma^2 / n$
- Hence, it is likely to be a very accurate estimate when $n$ is large

$$P(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha$$
Confidence Interval of $\mu$; $\sigma$ Known

\[ P\left( -z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \right) \]

\[ = P\left( \frac{\bar{X} - \sigma}{\sqrt{n}} z_{\alpha/2} < \mu < \frac{\sigma}{\sqrt{n}} z_{\alpha/2} + \bar{X} \right) = 1 - \alpha \]

- If $\bar{x}$ is the mean of random sample of size $n$ from a population with known variance $\sigma^2$, a $(1-\alpha)\times100\%$ confidence interval for $\mu$ is given by

\[ \bar{x} - \frac{\sigma}{\sqrt{n}} \frac{z_{\alpha}}{2} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} \frac{z_{\alpha}}{2} \]

This works well for $n \geq 30$ even if the population is not normal.
Example 9.2

- The average zinc concentration in 36 different locations is 2.6 gm/ml. Find the 95% and 90% confidence intervals for the mean zinc concentration. Assume $\sigma = 0.3$

- Solution:

\[
Z_{0.05/2} = Z_{0.025} = 1.96 \quad Z_{0.01/2} = Z_{0.005} = 2.575
\]

\[
\bar{X} - \frac{Z_{\alpha/2}\sigma}{\sqrt{n}} < \mu < \bar{X} + \frac{Z_{\alpha/2}\sigma}{\sqrt{n}}
\]

\[
2.6 - \frac{1.96(0.3)}{\sqrt{36}} < \mu < 2.6 + \frac{1.96(0.3)}{\sqrt{36}}
\]

\[
2.50 < \mu < 2.70
\]

\[
2.6 - \frac{2.575(0.3)}{\sqrt{36}} < \mu < 2.6 + \frac{2.575(0.3)}{\sqrt{36}}
\]

\[
2.47 < \mu < 2.73
\]
Estimation Error

• Since

\[ \bar{x} - \frac{\sigma}{\sqrt{n}} \frac{z_\alpha}{2} < \mu < \bar{x} + \frac{\sigma}{\sqrt{n}} \frac{z_\alpha}{2} \]

If \( \bar{x} \) is used as an estimate of \( \mu \), we can then be \((1-\alpha)\times100\%\) confidence that the error will not exceed \( \frac{\sigma \frac{z_{\alpha/2}}{\sqrt{n}}}{\sqrt{n}} \)

\[
\frac{\sigma \frac{z_{\alpha/2}}{\sqrt{n}}}{\sqrt{n}} = e \quad \Rightarrow \quad n = \left( \frac{\sigma \frac{z_{\alpha/2}}{e}}{e} \right)^2
\]

This specified the sample size.
Example 9.3

• How large a sample is required in Example 9.2 if we want to be 95% confident that our estimate of $\mu$ is off by less than 0.05?

• Solution: Population std dev given as 0.3. By Theorem 9.2:

$$n = \left( \frac{Z_{\alpha/2} \sigma}{e} \right)^2$$

$$n = \left( \frac{1.96(0.3)}{0.05} \right)^2 = 138.3 \approx 139$$
In case of unknown $\sigma$

- In real case, the variance is known...
- But, we have the sample standard deviation $S$
- Recall that the Student’s $t$-distribution is with $n-1$ degrees of freedom.

$$T = \frac{\overline{X} - \mu}{S / \sqrt{n}}$$

Figure 9.5 $P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha.$
Confidence Interval of $\mu$; $\sigma$ unknown

\[ P \left( -t_{\alpha/2} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2} \right) \]

Replace normal by t-distribution
Replace $\sigma$ by $S$

\[ = P \left( \bar{X} - \frac{S}{\sqrt{n}} t_{\alpha/2} < \mu < \bar{X} + \frac{S}{\sqrt{n}} t_{\alpha/2} \right) = 1 - \alpha \]

- If $\bar{x}$ and $s$ are the mean and std. dev. of a random sample of size $n$ from a normal population with unknown variance $\sigma^2$, a $(1-\alpha) \times 100\%$ confidence interval for $\mu$ is given by

\[ \bar{x} - \frac{s}{\sqrt{n}} t_{\alpha/2} < \mu < \bar{x} + \frac{s}{\sqrt{n}} t_{\alpha/2} \]

$t_{\alpha/2}$ is the $t$-value with $\nu=n-1$ degrees of freedom, leaving an area of $\alpha/2$ to the right

Prob. & Stat. Lecture09 - one-/two-sample estimation (cwliu@twins.ee.nctu.edu.tw)
Large Sample Confidence Interval

- If $n \geq 30$, one can use normal distribution instead of Student’s $t$-distribution, even when normality cannot be assumed.
  - This is reasonable because of central limit theorem
Example 9.4

- 7 similar containers are 9.8, 10.2, 10.4, 9.8, 10.0, 10.2, and 9.6 liters. Find a 95% confidence interval for the mean of all such containers, assuming an approximately normal distribution.

- Solution:
  \[ \bar{x} = 10.0 \quad s = 0.283 \]

Using Table A.4 (Student-t), find \( t_{0.025} = 2.447 \) for \( v = 6 \)

\[
\bar{x} - t_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right) < \mu < \bar{x} + t_{\alpha/2} \left( \frac{s}{\sqrt{n}} \right)
\]

\[
10 - 2.447 \left( \frac{0.283}{\sqrt{7}} \right) < \mu < 10 + 2.447 \left( \frac{0.283}{\sqrt{7}} \right)
\]

\[ 9.74 < \mu < 10.26 \]
Standard Error of a Point Estimate

The variance of \( \bar{X} \): \( \sigma^2_{\bar{X}} = \frac{\sigma^2}{n} \). \( \text{Point estimate} \)

Thus the standard deviation of \( \bar{X} \) or standard error of \( \bar{X} \): \( \sigma/\sqrt{n} \).

Confidence limit: \( \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \bar{x} \pm z_{\alpha/2} \text{s.e.}(\bar{x}). \)

Confidence limit with \( \sigma \) unknown: \( \bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}} = \bar{x} \pm t_{\alpha/2} \text{s.e.}(\bar{x}). \)

- Width of confidence intervals become shorter as the quality of the corresponding point estimate becomes better.
Prediction Interval

- Sometimes, some experimenters may also be interested in predicting the possible value of a future observation.
- Some customers may require a statement regarding the uncertainty of one single observation.
- The type of requirement is nicely fulfilled by the construction of a prediction interval.
- The variance of the random error in a new observation: $\sigma^2$

\[
Z = \frac{x_0 - \bar{x}}{\sqrt{\sigma^2 + \sigma_x^2 / n}} = \frac{x_0 - \bar{x}}{\sigma \sqrt{1 + 1/n}}
\]

Using $\Pr(z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

\[
\Rightarrow \bar{x} - z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}}
\]
Prediction Interval

• For a normal distribution of measurements with unknown mean $\mu$ and known variance $\sigma^2$, a $(1 - \alpha)100\%$ prediction interval of a future observation, $x_0$, is

$$\bar{x} - z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}}$$

where $z_{\alpha/2}$ is the $z$-value leaving an area of $\alpha/2$ to the right.

• Example 9.5: Due to the decreasing of interest rates, the First Citizens Bank received a lot of mortgage applications. A recent sample of 50 mortgage loans resulted in an average of $128,300. Assume a population standard deviation of $15,000. If a next customer called in for a mortgage loan application, find a 95% prediction interval on this customer’s loan amount.
Prediction Interval

\[ \bar{x} - z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + z_{\alpha/2} \sigma \sqrt{1 + \frac{1}{n}} \]

\[ \bar{x} = \$128,300, \, z_{0.025} = 1.96 \]

\[ 128300 - 1.96 \cdot 15000 \sqrt{1 + \frac{1}{50}} < x < 128300 + 1.96 \cdot 15000 \sqrt{1 + \frac{1}{50}} \]

\therefore \text{prediction interval} (\$98,607.46, \$157,992.54)

• Prediction interval provides a good estimate of the location of a future observation.
• The estimation of future observation is quite different from the estimation of sample mean.
Prediction Interval

• For a normal distribution of measurements with unknown mean $\mu$ and unknown variance $\sigma^2$, a $(1 - \alpha)100\%$ prediction interval of a future observation, $x_0$, is

$$\bar{x} - t_{\alpha/2} s \sqrt{1 + \frac{1}{n}} < x_0 < \bar{x} + t_{\alpha/2} s \sqrt{1 + \frac{1}{n}},$$

where $t_{\alpha/2}$ is the $t$-value with $\nu = n - 1$ degree-of-freedom, leaving an area of $\alpha/2$ to the right.

• Example 9.6: A meat inspector has randomly measured 30 packs of acclaimed 95% lean beef. The sample resulted in the mean 96.2% with the sample standard deviation of 0.8%. Find a 99% prediction interval for a new pack. Assume normality.
Prediction Interval

\[ v = 29, t_{0.005} = 2.756 \]

\[ 96.2 - 2.756 \cdot 0.8\sqrt{1 + \frac{1}{30}} < x < 96.2 + 2.756 \cdot 0.8\sqrt{1 + \frac{1}{30}} \]

\[ \therefore \text{prediction interval (93.96, 98.44)} \]

\[ x = t_\nu \frac{s}{\sqrt{n}} \]

- An observation is an outlier if it falls outside the prediction interval computed without inclusion of the questionable observation in the sample.
Tolerance Interval

- One method of establishing the desired bound is to determine a confidence interval on a fixed proportion of the measurements.
- For a normal distribution $N(\mu, \sigma)$, a bound that covers the middle 95% of the population of observations is then

$$\mu \pm 1.96\sigma$$

- This is called a tolerance interval, and indeed the coverage of 95% of measured observations is exact.
Tolerance Limits

• **Tolerance limits**: For a normal distribution of measurements with unknown mean $\mu$ and unknown variance $\sigma^2$, tolerance limits are given by $\bar{x} \pm ks$, where $k$ is determined so that one can assert with $100(1 - \gamma)\%$ confidence that the given limits contain at least the proportion $(1 - \alpha)$ of the measurements.

• **Example 9.7**: A machine is producing metal pieces that are cylindrical in shape. A sample of these pieces is taken and the diameters are found to be 1.01, 0.97, 1.03, 1.04, 0.99, 0.98, 0.99, 1.01, and 1.03 centimeters. Find the 99% tolerance limits that will contain 95% of the metal pieces produced by this machine, assuming an approximate normal distribution.

  - **Solution**

    $\bar{x} = 1.0056, \ s = 0.0246$

    Look up Table A.7, $n = 9, 1 - \gamma = 0.99$, and $1 - \alpha = 0.95 \Rightarrow k = 4.550$

    $: 99\%$ tolerance limits $: 1.0056 \pm 4.550 \cdot 0.0246$

    (cwliu@twins.ee.nctu.edu.tw)
### TABLE A.7* Tolerance Factors for Normal Distributions

| $\gamma = 0.05$ | $1 - \alpha$  |  |  |  |
|----------------|----------------|----------------|----------------|
| **n**          | **0.90**       | **0.95**       | **0.99**       |  |
| 2              | 32.019         | 37.674         | 48.430         |  |
| 3              | 8.380          | 9.916          | 12.861         |  |
| 4              | 5.369          | 6.370          | 8.299          |  |
| 5              | 4.275          | 5.079          | 6.634          |  |
| 6              | 3.712          | 4.414          | 5.775          |  |
| 7              | 3.369          | 4.007          | 5.248          |  |
| 8              | 3.136          | 3.732          | 4.891          |  |
| 9              | 2.967          | 3.532          | 4.631          |  |
| 10             | 2.839          | 3.379          | 4.433          |  |
| 11             | 2.737          | 3.259          | 4.277          |  |
| 12             | 2.655          | 3.162          | 4.150          |  |
| 13             | 2.587          | 3.081          | 4.044          |  |
| 14             | 2.529          | 3.012          | 3.955          |  |
| 15             | 2.480          | 2.954          | 3.878          |  |
| 16             | 2.437          | 2.903          | 3.812          |  |
| 17             | 2.400          | 2.858          | 3.754          |  |
| 18             | 2.366          | 2.819          | 3.702          |  |
| 19             | 2.337          | 2.784          | 3.656          |  |
| 20             | 2.310          | 2.752          | 3.615          |  |
| 25             | 2.208          | 2.631          | 3.457          |  |
| 30             | 2.140          | 9.549          | 3.350          |  |
| 35             | 2.090          | 2.490          | 3.272          |  |
| 40             | 2.052          | 2.445          | 3.213          |  |

| $\gamma = 0.01$ | $1 - \alpha$  |  |  |  |
|----------------|----------------|----------------|----------------|
| **n**          | **0.90**       | **0.95**       | **0.99**       |  |
| 2              | 160.193        | 188.491        | 242.300        |  |
| 3              | 18.930         | 22.401         | 29.055         |  |
| 4              | 9.398          | 11.150         | 14.527         |  |
| 5              | 6.612          | 7.855          | 10.260         |  |
| 6              | 5.337          | 6.345          | 8.301          |  |
| 7              | 4.613          | 5.488          | 7.187          |  |
| 8              | 4.147          | 4.936          | 6.468          |  |
| 9              | 3.822          | 4.550          | 5.966          |  |
| 10             | 3.582          | 4.265          | 5.594          |  |
| 11             | 3.397          | 4.045          | 5.308          |  |
| 12             | 3.250          | 3.870          | 5.079          |  |
| 13             | 3.130          | 3.727          | 4.893          |  |
| 14             | 3.029          | 3.608          | 4.737          |  |
| 15             | 2.945          | 3.507          | 4.605          |  |
| 16             | 2.872          | 3.421          | 4.492          |  |
| 17             | 2.808          | 3.345          | 4.393          |  |
| 18             | 2.753          | 3.279          | 4.307          |  |
| 19             | 2.703          | 3.221          | 4.230          |  |
| 20             | 2.659          | 3.168          | 4.161          |  |
| 25             | 2.494          | 2.972          | 3.904          |  |
| 30             | 2.385          | 2.841          | 3.733          |  |
| 35             | 2.306          | 2.748          | 3.611          |  |
| 40             | 2.247          | 2.677          | 3.518          |  |

Distinction Among Confidence Intervals, Prediction Intervals, and Tolerance Intervals

• Confidence intervals: population mean
• Tolerance limits: a tolerance interval must necessarily be longer than a confidence interval with the same degree of confidence.
• Prediction limits: determine a bound of a future observation value.
Estimating the Difference Between Two Means

- **Two populations:** $(\mu_1, \sigma_1), (\mu_2, \sigma_2)$
- **The statistics** $\bar{X}_1 - \bar{X}_2$ is a point estimator for $\mu_1 - \mu_2$

The sampling distribution of $\bar{X}_1 - \bar{X}_2$ will be approximated normally with mean $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$ and standard deviation $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

$\Rightarrow$ standard normal variable $Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$

$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$

$P(-z_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < z_{\alpha/2}) = 1 - \alpha$

- **Confidence interval**

A $(1 - \alpha)100\%$ confidence interval for mean $\mu_1 - \mu_2$ is given by

$$(\bar{X}_1 - \bar{X}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{X}_1 - \bar{X}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$
Example 9.8

- An experiment was conducted in which two types of engines, A and B, were compared. Gas mileage in miles per gallon was measured.
  - Fifty experiments were conducted using engine type A and 75 experiments were done for engine type B.
  - The average gas mileage for engine A was 36 miles per gallon and the average for machine B was 42 miles per gallon.
  - Find a 96% confidence interval on $\mu_B - \mu_A$, where $\mu_B$ and $\mu_A$ are population mean gas mileage for machines B and A, respectively.
  - Assume that the population standard deviations are 6 and 8 for machines A and B, respectively.

- Solution

\[
(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}
\]

\[
(42 - 36) - 2.05 \sqrt{\frac{64}{75} + \frac{36}{50}} < \mu_1 - \mu_2 < (42 - 36) + 2.05 \sqrt{\frac{64}{75} + \frac{36}{50}}
\]
Estimating the Difference Between Two Means with Unknown $\sigma$

- Variance unknown: If $\sigma_1^2$ and $\sigma_2^2$ are unknown, but $\sigma_1^2 = \sigma_2^2 = \sigma^2$, we obtain

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}}$$

Two chi-squared random variables: $\frac{(n_1 - 1)S_1^2}{\sigma_1^2}$ and $\frac{(n_2 - 1)S_2^2}{\sigma_2^2}$

Their sum: $V = \frac{(n_1 - 1)S_1^2}{\sigma_1^2} + \frac{(n_2 - 1)S_2^2}{\sigma_2^2} = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{\sigma^2}$

has a chi-squared distribution with $\nu = n_1 + n_2 - 2$ degrees of freedom
Estimating the Difference Between Two Means with Unknown $\sigma$

- From Theorem 8.5:
  \[ T = \frac{Z}{\sqrt{V / v}} \]
  \[ \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2[(1/n_1) + (1/n_2)]}} \frac{1}{\sqrt{\sqrt{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}}} \]
  \[ \frac{\sigma^2(n_1 + n_2 - 2)}{\sqrt{\sigma^2(n_1 + n_2 - 2)}} \]

- Pooled estimator of the unknown common variance $\sigma^2$:
  \[ S_P^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \]
  \[ \therefore T = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_P\sqrt{(1/n_1) + (1/n_2)}} \]
Estimating the Difference Between Two Means with Unknown $\sigma$

- \[ P(-t_{\alpha/2} < T < t_{\alpha/2}) = 1 - \alpha \]

\[
P\left[-t_{\alpha/2} < \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{S_p \sqrt{1/n_1 + 1/n_2}} < t_{\alpha/2}\right] = 1 - \alpha
\]

- Confidence interval for $\mu_1 - \mu_2$; $\sigma_1^2 = \sigma_2^2$ but unknown:

\[
(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}
\]

$s_p$ : the pooled estimate of the population standard deviation

$t_{\alpha/2}$ : the $t$-value with $\nu = n_1 + n_2 - 2$ degrees of freedom
Example 9.9

- Two independent sampling stations were chosen for the study of acid mine pollution.
  - For 12 monthly samples collected at the downstream station the species diversity index had a mean value $\bar{x}_1 = 3.11$ and a standard deviation $s_1 = 0.771$, while 10 monthly samples collected at the upstream station the species diversity index had a mean value $\bar{x}_2 = 2.04$ and a standard deviation $s_2 = 0.448$.
  - Find a 90% confidence interval for the difference between the population means for the two locations, assuming that the population are approximately normally distributed with equal variances.

- Solution

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = \frac{(12 - 1)0.771^2 + (10 - 1)0.448^2}{12 + 10 - 2} = 0.471 \Rightarrow S_p = 0.646$$

$$(3.11 - 2.04) - t_{0.05}(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}} < \mu_1 - \mu_2 < (3.11 - 2.04) + t_{0.05}(0.646)\sqrt{\frac{1}{12} + \frac{1}{10}}$$

$\therefore 0.593 < \mu_1 - \mu_2 < 1.547$
Estimating the Difference Between Two Means with Unequal Variances

• Unequal Variances

\[ T' = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{(S_1^2 / n_1) + (S_2^2 / n_2)}} \]

\[ \nu = \frac{(s_1^2 / n_1 + s_2^2 / n_2)^2}{[(s_1^2 / n_1)^2 / (n_1 - 1)] + [(s_2^2 / n_2)^2 / (n_2 - 1)]} \] (degrees of freedom)

\[ P(-t_{\alpha/2} < T' < t_{\alpha/2}) \approx 1 - \alpha \]

• Confidence interval for \( \mu_1 - \mu_2; \) \( \sigma_1^2 \neq \sigma_2^2 \) and unknown:

\[ (\bar{x}_1 - \bar{x}_2) - t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + t_{\alpha/2} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \]
Example 9.10

- Zinc is measured in milligrams per liter. 15 samples were collected from station 1 had an average zinc content of 3.84 milligrams per liter and a standard deviation of 3.07 milligrams per liter, while the 12 samples from station 2 had an average zinc content of 1.49 milligrams per liter and a standard deviation of 0.80 milligrams per liter. Find a 95% confidence interval for the difference in the true average zinc contents at these two stations, assuming that the observations came from normal population with different variance.

- Solution

\[
v = \frac{(3.07^2/15 + 0.80^2/12)^2}{[(3.07^2/15)^2/14] + [(0.80^2/12)^2/11]} = 16.3 \approx 16
\]

For 95% confidence interval \( \Rightarrow t_{0.025} = 2.120 \) for \( v = 16 \)

\[
(3.84 - 1.49) - (2.120)\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}} < \mu_1 - \mu_2 < (3.84 - 1.49) + (2.120)\sqrt{\frac{3.07^2}{15} + \frac{0.80^2}{12}}
\]

\[
0.60 < \mu_1 - \mu_2 < 4.10
\]
Paired Observations

- Consider that the samples are not independent and the variances of the two populations are not necessarily equal.
- If $\bar{d}$ and $s_d$ are the mean and standard deviation of the normally distributed differences of $n$ random pairs of measurements, a $(1-a)100\%$ confidence interval for $\mu_D = \mu_1 - \mu_2$ is

$$\bar{d} - t_{\alpha/2} \frac{s_d}{\sqrt{n}} < \mu_D < \bar{d} + t_{\alpha/2} \frac{s_d}{\sqrt{n}}$$

where $t_{\alpha/2}$ is the $t$-value with $\nu = n - 1$ degrees of freedom, leaving an area of $\alpha/2$ to the right.

- Example 9.11: For a study of dioxin, find a 95% confidence interval for $\mu_1 - \mu_2$, where $\mu_1$ and $\mu_2$ represent the true mean TCDD in plasma and in fat tissue, respectively. Assume the distribution of the differences to be approximately normal.
Paired Observations

<table>
<thead>
<tr>
<th>Veteran</th>
<th>TCDD levels In plasma</th>
<th>TCDD levels In fat tissue</th>
<th>$d_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.5</td>
<td>4.9</td>
<td>-2.4</td>
</tr>
<tr>
<td>2</td>
<td>3.1</td>
<td>5.9</td>
<td>-2.8</td>
</tr>
<tr>
<td>3</td>
<td>2.1</td>
<td>4.4</td>
<td>-2.3</td>
</tr>
<tr>
<td>4</td>
<td>3.5</td>
<td>6.9</td>
<td>-3.4</td>
</tr>
<tr>
<td>5</td>
<td>3.1</td>
<td>7.0</td>
<td>-3.9</td>
</tr>
<tr>
<td>6</td>
<td>1.8</td>
<td>4.2</td>
<td>-2.4</td>
</tr>
<tr>
<td>7</td>
<td>6.0</td>
<td>10.0</td>
<td>-4.0</td>
</tr>
<tr>
<td>8</td>
<td>3.0</td>
<td>5.5</td>
<td>-2.5</td>
</tr>
<tr>
<td>9</td>
<td>36.0</td>
<td>41.0</td>
<td>-5.0</td>
</tr>
<tr>
<td>10</td>
<td>4.7</td>
<td>4.4</td>
<td>0.3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Veteran</th>
<th>TCDD levels In plasma</th>
<th>TCDD levels In fat tissue</th>
<th>$d_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>6.9</td>
<td>7.0</td>
<td>-0.1</td>
</tr>
<tr>
<td>12</td>
<td>3.3</td>
<td>2.9</td>
<td>0.4</td>
</tr>
<tr>
<td>13</td>
<td>4.6</td>
<td>4.6</td>
<td>0.0</td>
</tr>
<tr>
<td>14</td>
<td>1.6</td>
<td>1.4</td>
<td>0.2</td>
</tr>
<tr>
<td>15</td>
<td>7.2</td>
<td>7.7</td>
<td>-0.5</td>
</tr>
<tr>
<td>16</td>
<td>1.8</td>
<td>1.1</td>
<td>0.7</td>
</tr>
<tr>
<td>17</td>
<td>20.0</td>
<td>11.0</td>
<td>9.0</td>
</tr>
<tr>
<td>18</td>
<td>2.0</td>
<td>2.5</td>
<td>-0.5</td>
</tr>
<tr>
<td>19</td>
<td>2.5</td>
<td>2.3</td>
<td>0.2</td>
</tr>
<tr>
<td>20</td>
<td>4.1</td>
<td>2.5</td>
<td>1.6</td>
</tr>
</tbody>
</table>


- **Solution**
  \[
  \bar{d} = -0.87, \ t_{0.025} = 2.093 \ (v = 20 - 1 = 19), \ s_d = \sqrt{\frac{\sum (d_i - \bar{d})^2}{n - 1}} = \sqrt{\frac{168.4220}{19}} = 2.9773
  \]
  \[
  -0.87 - \left( 2.093 \right) \frac{2.9773}{\sqrt{20}} < \mu_D < -0.87 + \left( 2.093 \right) \frac{2.9773}{\sqrt{20}}
  \]
  \[
  \Rightarrow -2.2634 < \mu_D < 0.5234 \ : \text{no significat difference}
  \]
Single Sample: Estimating a Proportion

- A point estimator of the proportion $p$ in a binomial experiment: $\hat{P} = X/n$

\[
\mu_{\hat{p}} = E(\hat{P}) = E\left(\frac{X}{n}\right) = \frac{np}{n} = p
\]

\[
\sigma_{\hat{p}}^2 = \sigma_x^2/n^2 = \frac{nq}{n^2} = \frac{pq}{n}
\]

\[
P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha, Z = \frac{\hat{P} - p}{\sqrt{pq/n}}
\]

\[
P(-z_{\alpha/2} < \frac{\hat{P} - p}{\sqrt{pq/n}} < z_{\alpha/2}) = 1 - \alpha
\]

\[
P(\hat{P} - z_{\alpha/2}\sqrt{\frac{pq}{n}} < \mu < \hat{P} + z_{\alpha/2}\sqrt{\frac{pq}{n}}) = 1 - \alpha
\]

\[
P(\hat{P} - z_{\alpha/2}\sqrt{\frac{\hat{pq}}{n}} < \mu < \hat{P} + z_{\alpha/2}\sqrt{\frac{\hat{pq}}{n}}) \approx 1 - \alpha \text{ (when } n \text{ is large } \hat{p} = x/n \approx p)
\]
Single Sample: Estimating a Proportion

- If $\hat{p}$ is the proportion of successes in a random sample of size $n$, and $\hat{q} = 1 - \hat{p}$ an approximate $(1-\alpha)100\%$ confidence interval for the binomial parameter $p$ is given by

$$\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

where $z_{\alpha/2}$ is the z-value with leaving an area of $\alpha/2$ to the right.

- $n\hat{p}, n\hat{q} \geq 5$. 
Example 9.12

- In a random of \( n = 500 \) families owning television sets in the city of Hamilton, Canada, it is found that \( x = 340 \) subscribed to HBO. Find a 95% confidence interval for the actual proportion of families in the city who subscribe to HBO.

- Solution

\[
\hat{p} - z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} < p < \hat{p} + z_{\alpha/2} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}
\]

\[
\frac{340}{500} - 1.96 \sqrt{\frac{0.68 \times 0.32}{500}} < p < \frac{340}{500} + 1.96 \sqrt{\frac{0.68 \times 0.32}{500}}
\]

\[
0.64 < p < 0.72
\]
Theorems 9.3 & 9.4

• If \( \hat{p} \) is used as an estimate of \( p \), we can be \((1 - \alpha)100\%\) confident that the error will not exceed \( z_{\alpha/2} \sqrt{\hat{p}\hat{q}/n} \).

![Figure 9.6 Error in estimating \( p \) by \( \hat{p} \).](image)

• If \( \hat{p} \) is used as an estimate of \( p \), we can be \((1 - \alpha)100\%\) confident that the error will be less than a specified amount \( e \) when the sample size is approximately

\[
n = \frac{z_{\alpha/2}^2 \hat{p}\hat{q}}{e^2}.
\]
Examples

• Example 9.13: How large a sample is required in Example 9.12 if we want to be 95% confident that our estimate of $p$ is within 0.02?

  - Solution

\[ n = \frac{z_{\alpha/2}^2 \hat{p} \hat{q}}{e^2} = \frac{(1.96)^2 (0.68)(0.32)}{(0.02)^2} = 2090. \]

\[ n = \frac{(1.96)^2}{4(0.02)^2} = 2401. \]
Theorem 9.5

- Upper bound of $n$ is $\frac{z_{\alpha/2}^2}{4e^2}$, $\therefore \hat{p}\hat{q} = \hat{p}(1-\hat{p}) = \frac{1}{4} - (\hat{p} - \frac{1}{2})^2$

- If $\hat{p}$ is used as an estimate of $p$, we can be at least ($1-\alpha$)100% confident that the error will not exceed a specified amount $e$ when the sample size is approximately $n = \frac{z_{\alpha/2}^2}{4e^2}$.

- Example 9.14: How large a sample is required in Example 9.12 if we want to be at least 95% confident that our estimate of $p$ is within 0.02?
  - Solution
    $$n = \frac{z_{\alpha/2}^2}{4e^2} = \frac{(1.96)^2}{4 \cdot (0.02)^2} = 2401.$$
Estimating the Difference Between Two Proportions

- $p_1$ might be the proportion of smokers with lung cancer and $p_2$ the proportion of non-smokers with lung cancer.

- The sampling distribution of $\hat{P}_1 - \hat{P}_2$ will be approximated normally with mean $\mu_{\hat{P}_1 - \hat{P}_2} = p_1 - p_2$ and variance $\sigma^2_{\hat{P}_1 - \hat{P}_2} = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$

\[ \Rightarrow \text{standard normal variable } Z = \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} \]

\[ P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \]

\[ P\left[ -z_{\alpha/2} < \frac{(\hat{P}_1 - \hat{P}_2) - (p_1 - p_2)}{\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}} < z_{\alpha/2} \right] = 1 - \alpha \]
Estimating the Difference Between Two Proportions

• Large-sample confidence interval for $p_1 - p_2$:
If $\hat{p}_1$ and $\hat{p}_2$ are the proportion of successes in a random sample of size $n_1$ and $n_2$, $\hat{q}_1 = 1 - \hat{p}_1$ and $\hat{q}_2 = 1 - \hat{p}_2$, an approximate $(1-\alpha)100\%$ confidence interval for the difference of two binomial parameters $p_1 - p_2$ is given by

$$(\hat{p}_1 - \hat{p}_2) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < (\hat{p}_1 - \hat{p}_2) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$$

where $z_{\alpha/2}$ is the $z$-value leaving an area of $\alpha/2$ to the right.

• $n_1 \hat{p}_1, n_1 \hat{q}_1, n_2 \hat{p}_2, n_2 \hat{q}_2 \geq 5$
Example 9.15

- A certain change in a process for manufacture of component parts is being considered.
  - Sample are taken using both the existing and new procedure so as to determine if the new process results in an improvement.
  - If 75 of 1500 items from the existing procedure were found to be defective and 80 of 2000 items from the new procedure were found to be defective.
  - Find a 90% confidence interval for the true difference in the fraction of defectives between the existing and the new process.

- Solution

\[
\left( \hat{p}_1 - \hat{p}_2 \right) - z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}} < p_1 - p_2 < \left( \hat{p}_1 - \hat{p}_2 \right) + z_{\alpha/2} \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}
\]

\[
\hat{p}_1 = 75/1500 = 0.05, \hat{p}_2 = 80/2000 = 0.04, \hat{p}_1 - \hat{p}_2 = 0.05 - 0.04 = 0.01
\]

\[
0.01 - 1.645 \sqrt{\frac{0.05 \cdot 0.95}{1500} + \frac{0.04 \cdot 0.96}{2000}} < p_1 - p_2 < 0.01 + 1.645 \sqrt{\frac{0.05 \cdot 0.95}{1500} + \frac{0.04 \cdot 0.96}{2000}}
\]

\[-0.0017 < p_1 - p_2 < 0.0217\]
Estimating Variance

- If a sample size of \( n \) is drawn from a normal population with variance \( \sigma^2 \), and the sample variance \( s^2 \) is computed, then \( s^2 \) is the point estimate of \( \sigma^2 \).
- The statistic \( S^2 \) is called an estimator of \( \sigma^2 \).
- An interval estimator of \( \sigma^2 \)

\[
P(\chi^2_{1-\alpha/2} < X^2 = \frac{(n-1)S^2}{\sigma^2} < \chi^2_{\alpha/2}) = 1 - \alpha
\]

Chi-squared distribution with \( n-1 \) degrees of freedom

**Figure 9.7** \( P(\chi^2_{1-\alpha/2} < X^2 < \chi^2_{\alpha/2}) = 1 - \alpha \).
Confidence Interval for $\sigma^2$

- If $s^2$ is the variance of a random sample of size $n$ from a normal population, a $(1-\alpha)\times100\%$ confidence interval for $\sigma^2$ is
  \[
  \frac{(n-1)s^2}{\chi^2_{\alpha/2}} < \sigma^2 < \frac{(n-1)\hat{s}^2}{\chi^2_{1-\alpha/2}}
  \]
  where $\chi^2_{\alpha/2}$ and $\chi^2_{1-\alpha/2}$ are $\chi^2$-values with $\nu=n-1$ degrees of freedom, leaving areas of $\alpha/2$ and $1-\alpha/2$, respectively, to the right.

- Example 9.16: 10 package of grass seed distributed by a certain company: 46.4, 46.1, 45.8, 46.1, 45.9, 45.8, 46.9, 45.2, and 46.0 (weights in decagrams). Find a 95% confidence interval for the variance of all such packages of grass seed distributed by this company, assuming a normal distribution.

- Solution:
  \[
  s^2 = 0.286. \text{ For 95\% confidence interval, } \alpha = 0.05. \\
  \]
  Using A.5, with $\nu = 9$: $\chi^2_{0.025} = 19.023$, $\chi^2_{0.975} = 2.700$
  \[
  \frac{9(0.286)}{19.023} < \sigma^2 < \frac{9(0.286)}{2.7}
  \]
Estimating the Ratio of Two $\sigma^2$'s

- The point estimate of the ratio of two population variances $\sigma_1^2/\sigma_2^2$ is given by the ratio $s_1^2/s_2^2$ of the sample variances.
- The statistic $S_1^2/S_2^2$ is the estimator of $\sigma_1^2/\sigma_2^2$.
- If we have normal distributions, then

$$P\left[f_{1-\alpha/2}(v_1, v_2) < F = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2} < f_{\alpha/2}(v_1, v_2)\right] = 1 - \alpha$$

F-distribution with $v_1$ and $v_2$ degrees of freedom.

**Figure 9.8** $P[f_{1-\alpha/2}(v_1, v_2) < F < f_{\alpha/2}(v_1, v_2)] = 1 - \alpha$. 

9-48
Confidence Interval for $\sigma_1^2/\sigma_2^2$

- If $s_1^2$ and $s_2^2$ are the variances of independent samples of size $n_1$ and $n_2$, respectively, from normal populations, a $(1-\alpha)\times100\%$ confidence interval for $\sigma_1^2/\sigma_2^2$ is

$$\frac{s_1^2}{s_2^2} \frac{1}{f_{\alpha/2}(v_1, v_2)} < \frac{\sigma_1^2}{\sigma_2^2} < \frac{s_1^2}{s_2^2} f_{\alpha/2}(v_2, v_1)$$

where $f_{\alpha/2}(v_1, v_2)$ is $f$-values with $v_1=\nu_1-1$ and $v_2=\nu_2-1$ degrees of freedom, leaving areas of $\alpha/2$ to the right, and $f_{\alpha/2}(v_2, v_1)$ is a similar $f$-value with $v_2=\nu_2-1$ and $v_1=\nu_1-1$ degrees of freedom.