Basic Statistics and Probability

Chapter 7:
Inferences Based on a Single Sample:
Estimation with Confidence Intervals

- Confidence Interval
- Confidence Interval for a Population Mean
- Confidence Interval for the Mean of a Normal Population
- Confidence Interval for a Population Proportion
- Determining the Sample Size
Confidence Interval

- In Ch. 6 we saw that a point estimator of a target population parameter $\theta$ is a specific quantity $\hat{\theta}$ estimating $\theta$ and calculated from the sample $X_1, \ldots, X_n$.
- A point estimate will “always” carry an error. We seek to measure the uncertainty inherent to the point estimator.
- An interval estimator or confidence interval for the target population parameter $\theta$ is a whole interval of values estimating $\theta$ and calculated from the sample $X_1, \ldots, X_n$.
- Interval estimation gives an interval containing the parameter $\theta$ with a predetermined high confidence level.
- The confidence level $1 - \alpha$ is a measure of our degree of certainty that $\theta$ will be in the interval.
- We will talk about, say, 90%, 95%, 99% confidence intervals (intervals with a confidence level of 90%, 95%, 99% or confidence coefficient of 0.9, 0.95, 0.99).
Interpretation of the confidence level

Suppose we observe 100 samples of size \( n \) from a r.v. \( X \) whose distribution depends on a parameter \( \theta \). Then we construct the corresponding 100 confidence intervals for \( \theta \) with confidence level \( 1 - \alpha \), \( \text{CI}_{1-\alpha}(\theta) \).

Sample 1 \( \rightarrow \) \( \text{CI}_{1-\alpha}^{(1)}(\theta) \)

Sample 2 \( \rightarrow \) \( \text{CI}_{1-\alpha}^{(2)}(\theta) \)

\[ \vdots \]

Sample 100 \( \rightarrow \) \( \text{CI}_{1-\alpha}^{(100)}(\theta) \)

The parameter \( \theta \) will be in approximately \((1 - \alpha)100\) of them.
Confidence Interval for a Population Mean

Let $X$ be a random variable with mean $\mu$ and standard deviation $\sigma$. For a sample of $X$ with sample mean $\bar{X}$, by the CLT we know that, for a large sample size $n$, the approximate distribution of the $z$-statistic

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad \text{if } \sigma \text{ unknown} \quad \sim \quad \frac{\bar{X} - \mu}{s/\sqrt{n}}.$$

is $N(0, 1)$.

Then, for large $n$ ($n \geq 20$), a confidence interval for $\mu$ at the confidence level $1 - \alpha$ is

$$CI_{1-\alpha}(\mu) = \left( \bar{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}} \right)$$
\( z_\alpha \) is that value leaving to its right an area equal to \( \alpha \) in the N(0,1) density:

Check on your own that:

\[
\begin{align*}
z_{0.05} &= 1.645 \\
z_{0.01} &= 2.33 \\
z_{0.005} &= 2.575
\end{align*}
\]
Exercise in McClave & Sincich: Latex allergy in health care workers. Health care workers who use latex gloves with glove powder may develop a latex allergy. Symptoms of a latex allergy include conjunctivitis, hand eczema, nasal congestion, a skin rash, and shortness of breath. Each in a sample of 46 hospital employees who were diagnosed with latex allergy reported on their exposure to latex gloves (Current Allergy & Clinical Immunology, Mar. 2004). Summary statistics for the number of latex gloves used per week are $\bar{x} = 19.3$ and $s = 11.9$.

a. Give a point estimate for the average number of latex gloves used per week by all health care workers with a latex allergy.

b. Form a 95% confidence interval for the average number of latex gloves used per week by all health care workers with a latex allergy.
Confidence Interval for the Mean of a Normal Population

If we know that the variable $X$ follows a normal distribution with mean $\mu$ (unknown) and standard deviation $\sigma$ (unknown), then it is possible to compute confidence intervals for the population mean $\mu$ for any sample size $n$ (even small ones).

To this end, we introduce a new probability distribution, called Student’s $t$.

The $t$ distribution is a continuous one, with a density which is symmetric with respecto to 0, but with tails heavier than those of the standard normal.

The $t$ distribution depends on a parameter called the degrees of freedom (d.f.). A $t$ distribution with, say, 3 d.f. is denoted by $t_3$.

As the d.f. increase the tails of the $t_{df}$ get lighter. When d.f. $\geq 30$, the $t_{df}$ can be approximated by a N(0,1).
### Table III  Critical Values of t

<table>
<thead>
<tr>
<th>Degrees of Freedom</th>
<th>$t_{.100}$</th>
<th>$t_{.050}$</th>
<th>$t_{.025}$</th>
<th>$t_{.010}$</th>
<th>$t_{.005}$</th>
<th>$t_{.001}$</th>
<th>$t_{.0005}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.078</td>
<td>6.314</td>
<td>12.706</td>
<td>31.821</td>
<td>63.657</td>
<td>318.31</td>
<td>636.62</td>
</tr>
<tr>
<td>2</td>
<td>1.886</td>
<td>2.920</td>
<td>4.303</td>
<td>6.965</td>
<td>9.925</td>
<td>22.326</td>
<td>31.598</td>
</tr>
<tr>
<td>3</td>
<td>1.638</td>
<td>2.353</td>
<td>3.182</td>
<td>4.341</td>
<td>5.841</td>
<td>10.213</td>
<td>12.924</td>
</tr>
<tr>
<td>4</td>
<td>1.533</td>
<td>2.132</td>
<td>2.776</td>
<td>3.747</td>
<td>4.604</td>
<td>7.173</td>
<td>8.610</td>
</tr>
<tr>
<td>5</td>
<td>1.476</td>
<td>2.015</td>
<td>2.571</td>
<td>3.365</td>
<td>4.032</td>
<td>5.893</td>
<td>6.869</td>
</tr>
<tr>
<td>6</td>
<td>1.440</td>
<td>1.943</td>
<td>2.447</td>
<td>3.143</td>
<td>3.707</td>
<td>5.208</td>
<td>5.950</td>
</tr>
<tr>
<td>7</td>
<td>1.415</td>
<td>1.895</td>
<td>2.365</td>
<td>2.998</td>
<td>3.499</td>
<td>4.785</td>
<td>5.408</td>
</tr>
<tr>
<td>8</td>
<td>1.397</td>
<td>1.860</td>
<td>2.306</td>
<td>2.896</td>
<td>3.355</td>
<td>4.501</td>
<td>5.041</td>
</tr>
<tr>
<td>9</td>
<td>1.383</td>
<td>1.833</td>
<td>2.262</td>
<td>2.821</td>
<td>3.250</td>
<td>4.297</td>
<td>4.781</td>
</tr>
<tr>
<td>10</td>
<td>1.372</td>
<td>1.812</td>
<td>2.228</td>
<td>2.764</td>
<td>3.169</td>
<td>4.144</td>
<td>4.587</td>
</tr>
<tr>
<td>11</td>
<td>1.363</td>
<td>1.796</td>
<td>2.201</td>
<td>2.718</td>
<td>3.106</td>
<td>4.025</td>
<td>4.437</td>
</tr>
<tr>
<td>12</td>
<td>1.356</td>
<td>1.782</td>
<td>2.179</td>
<td>2.681</td>
<td>3.055</td>
<td>3.930</td>
<td>4.318</td>
</tr>
<tr>
<td>13</td>
<td>1.350</td>
<td>1.771</td>
<td>2.160</td>
<td>2.650</td>
<td>3.012</td>
<td>3.852</td>
<td>4.221</td>
</tr>
<tr>
<td>14</td>
<td>1.345</td>
<td>1.761</td>
<td>2.145</td>
<td>2.624</td>
<td>2.977</td>
<td>3.787</td>
<td>4.140</td>
</tr>
<tr>
<td>15</td>
<td>1.341</td>
<td>1.753</td>
<td>2.131</td>
<td>2.602</td>
<td>2.947</td>
<td>3.733</td>
<td>4.073</td>
</tr>
<tr>
<td>16</td>
<td>1.337</td>
<td>1.746</td>
<td>2.120</td>
<td>2.583</td>
<td>2.921</td>
<td>3.686</td>
<td>4.015</td>
</tr>
<tr>
<td>17</td>
<td>1.333</td>
<td>1.740</td>
<td>2.110</td>
<td>2.567</td>
<td>2.898</td>
<td>3.646</td>
<td>3.965</td>
</tr>
<tr>
<td>18</td>
<td>1.330</td>
<td>1.734</td>
<td>2.101</td>
<td>2.552</td>
<td>2.878</td>
<td>3.610</td>
<td>3.922</td>
</tr>
<tr>
<td>19</td>
<td>1.328</td>
<td>1.729</td>
<td>2.093</td>
<td>2.539</td>
<td>2.861</td>
<td>3.579</td>
<td>3.883</td>
</tr>
<tr>
<td>20</td>
<td>1.325</td>
<td>1.725</td>
<td>2.086</td>
<td>2.528</td>
<td>2.845</td>
<td>3.552</td>
<td>3.850</td>
</tr>
<tr>
<td>21</td>
<td>1.323</td>
<td>1.721</td>
<td>2.080</td>
<td>2.518</td>
<td>2.831</td>
<td>3.527</td>
<td>3.819</td>
</tr>
<tr>
<td>22</td>
<td>1.321</td>
<td>1.717</td>
<td>2.074</td>
<td>2.508</td>
<td>2.819</td>
<td>3.505</td>
<td>3.792</td>
</tr>
<tr>
<td>23</td>
<td>1.319</td>
<td>1.714</td>
<td>2.069</td>
<td>2.500</td>
<td>2.807</td>
<td>3.485</td>
<td>3.767</td>
</tr>
<tr>
<td>24</td>
<td>1.318</td>
<td>1.711</td>
<td>2.064</td>
<td>2.492</td>
<td>2.797</td>
<td>3.467</td>
<td>3.745</td>
</tr>
<tr>
<td>25</td>
<td>1.316</td>
<td>1.708</td>
<td>2.060</td>
<td>2.485</td>
<td>2.787</td>
<td>3.450</td>
<td>3.725</td>
</tr>
<tr>
<td>26</td>
<td>1.315</td>
<td>1.706</td>
<td>2.056</td>
<td>2.479</td>
<td>2.779</td>
<td>3.435</td>
<td>3.707</td>
</tr>
<tr>
<td>27</td>
<td>1.314</td>
<td>1.703</td>
<td>2.052</td>
<td>2.473</td>
<td>2.771</td>
<td>3.421</td>
<td>3.690</td>
</tr>
<tr>
<td>28</td>
<td>1.313</td>
<td>1.701</td>
<td>2.048</td>
<td>2.467</td>
<td>2.763</td>
<td>3.408</td>
<td>3.674</td>
</tr>
<tr>
<td>29</td>
<td>1.311</td>
<td>1.699</td>
<td>2.045</td>
<td>2.462</td>
<td>2.756</td>
<td>3.396</td>
<td>3.659</td>
</tr>
<tr>
<td>30</td>
<td>1.310</td>
<td>1.697</td>
<td>2.042</td>
<td>2.457</td>
<td>2.750</td>
<td>3.385</td>
<td>3.646</td>
</tr>
<tr>
<td>40</td>
<td>1.303</td>
<td>1.684</td>
<td>2.021</td>
<td>2.423</td>
<td>2.704</td>
<td>3.307</td>
<td>3.551</td>
</tr>
<tr>
<td>60</td>
<td>1.296</td>
<td>1.671</td>
<td>2.000</td>
<td>2.390</td>
<td>2.660</td>
<td>3.232</td>
<td>3.460</td>
</tr>
<tr>
<td>120</td>
<td>1.289</td>
<td>1.658</td>
<td>1.980</td>
<td>2.358</td>
<td>2.617</td>
<td>3.160</td>
<td>3.373</td>
</tr>
</tbody>
</table>
\( t_{df;\alpha} \) is the value leaving to its right an area \( \alpha \) in the \( t_{df} \) density:

If \( df \geq 30 \), then \( t_{df;\alpha} \approx z_{\alpha} \).

Check on your own that:

\[
t_{10;0.05} = 1.812 \quad t_{4;0.01} = 3.747 \quad t_{5;0.005} = 4.032
\]
If $X \sim N(\mu, \sigma)$ and $X_1, \ldots, X_n$ is a sample from $X$ with sample mean $\bar{X}$, then the $t$-statistic

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

follows a $t_{n-1}$ distribution. This result holds for any $n$.

Then a confidence interval for the mean $\mu$ of a normal population at the confidence level $1 - \alpha$ is

$$CI_{1-\alpha}(\mu) = \left( \bar{x} \pm t_{n-1;\alpha/2} \frac{s}{\sqrt{n}} \right)$$
Exercise in McClave & Sincich: Radon exposure in Egyptian tombs. Many ancient Egyptian tombs were cut from limestone rock that contained uranium. Since most tombs are not well-ventilated, guards, tour guides, and visitors may be exposed to deadly radon gas. In Radiation Protection Dosimetry (Dec. 2010), a study of radon exposure in tombs in the Valley of Kings, Luxor, Egypt (recently opened for public tours), was conducted. The radon levels – measured in becquerels per cubic meter (Bq/m³) – in the inner chambers of a sample of 12 tombs were determined. For this data, assume that $\bar{x} = 3,643$ Bq/m³ and $s = 1,187$ Bq/m³. Use this information to estimate, with 95% confidence, the true mean level of radon exposure in tombs in the Valley of Kings.
Confidence Interval for a Population Proportion

Let $X$ be a Bernoulli($p$) r.v.

$$X = \begin{cases} 
1 & \text{with probability } p \\
0 & \text{with probability } 1 - p 
\end{cases}$$

For a sample of size $n$ from $X$ we compute its sample mean, $\hat{p} = \bar{X}$. By the CLT, for large $n$, the $z$-statistic

$$Z = \frac{\hat{p} - p}{\sqrt{\hat{p}(1 - \hat{p})/n}}$$

is approximately $N(0, 1)$.

Then, for large $n$ ($n \geq 20$) and $0.1 \leq p \leq 0.9$, a confidence interval for $p$ at the confidence level $1 - \alpha$ is

$$\text{CI}_{1-\alpha}(p) = \left( \hat{p} \pm \frac{z_{\alpha/2}}{\sqrt{n}} \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}} \right).$$
Exercise in McClave & Sincich: Paying for music downloads.
If you use the Internet, have you ever paid to access or download music? This was one of the questions of interest in a recent *Pew Internet and American Life Project Survey* (Oct. 2010). Telephone interviews were conducted on a representative sample of 1,003 adults living in the United States. For this sample, 506 adults stated that they have paid to download music.

**a.** Use the survey information to find a point estimate for the true proportion of U.S. adults who have paid to download music.

**b.** Find an interval estimate for the proportion, part a. Use a 90% confidence interval.

**c.** Give a practical interpretation of the interval, part b. Your answer should begin with “We are 90% confident ...”
Determining the Sample Size

The appropriate sample size \( n \) for making an inference about a population mean or proportion depends on the desired certainty.

The reliability of a confidence interval for the population mean \( \mu \) or proportion \( p \) is given by the sampling error \( SE \) within which we want to estimate \( \mu \) or \( p \) with that confidence level:

\[
SE = \text{Half-width of the confidence interval.}
\]

**Example (Latex allergy in health care workers):**

Suppose we want to estimate \( \mu \), the expected number of latex gloves used per week by health care workers with a latex allergy, with a confidence level of 95% and a sampling error \( SE \) of 1. Which is the minimum required sample size?
Example (Radon exposure in Egyptian tombs):
What sample size should they use if the researchers want to estimate the mean level of radon exposure in those tombs to within 300 Bq/m$^3$ of its true value?

Example (Paying for music downloads):
How many telephone interviews should be conducted in order to estimate the proportion $p$ of U.S. adults who have paid to download music to within 0.01 with 90% confidence?