Statistics: Estimation

1 Reminder: Continuous variables

2 Point estimation

3 Confidence intervals
Reminder: Continuous variables

**Probability density function of \( x \)**

\[
P(a \leq x \leq b) = \int_{a}^{b} f(x) \, dx \quad \int_{\text{min}}^{\text{max}} f(x) \, dx = 1
\]

**Standard normal distribution \( \mathcal{N}(0, 1) \)**

- approx. 68\% of values at \( \mu \pm 1\sigma \)
- approx. 95\% of values at \( \mu \pm 2\sigma \)
- approx. 99\% of values at \( \mu \pm 3\sigma \)

**Standardized score**

\[
Z = \frac{x - \mu}{\sigma}
\]
The main puzzle

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Notation</th>
<th>Sample</th>
<th>Population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>( \bar{X} )</td>
<td>( \mu )</td>
<td></td>
</tr>
<tr>
<td>Standard deviation</td>
<td>( s )</td>
<td>( \sigma )</td>
<td></td>
</tr>
</tbody>
</table>

The main solution: the properties of the standard normal distribution allow for statistical \textit{inference}: the \textit{estimation}, at a certain level of \textit{confidence}, of the unobserved \textit{population} parameters, using observed \textit{sample} parameters.
Point estimation

Sample definitions

■ the population mean $\mu$ is a population parameter
■ the sample mean $\bar{X}$ is a point estimate of $\mu$
■ we know the sample $n$ and its mean $\bar{X}$, but we do not know $\mu$
  and might not know the true population $N$

Sampling error

■ sampling variation causes $\bar{X}$ to vary
■ the values of $\bar{X}$ form a sampling distribution
■ its standard deviation $\frac{\sigma}{\sqrt{n}}$ is the standard error of the mean (SEM), which is estimated from the sample
## CLT and LLN

### Central Limit Theorem (CLT)

For ‘iid’ (independent and identically distributed) random variables $X_1, X_2, ..., X_n$, the sampling distribution of the mean approximates a normal distribution as $n > 30$ increases.

$$
\sqrt{N}\left(\frac{1}{N} \sum_{i=1}^{N} \bar{X}_i - \mu\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2)
$$

### Law of Large Numbers (LLN)

$$
\frac{X_1 + X_2 + ... + X_n}{n} = \mu
$$
Here, we present a useful rule of thumb for the probability of falling within 1, 2, and 3 standard deviations of the mean in the normal distribution. This will be useful in a wide range of practical settings, especially when trying to make a quick estimate without a calculator or Z table.

Figure 3.9: Probabilities for falling within 1, 2, and 3 standard deviations of the mean in a normal distribution.

Exercise 3.22
Use the Z table to confirm that about 68%, 95%, and 99.7% of observations fall within 1, 2, and 3 standard deviations of the mean in the normal distribution, respectively. For instance, first find the area that falls between $Z = 1$ and $Z = 1$, which should have an area of about 0.68. Similarly, there should be an area of about 0.95 between $Z = 2$ and $Z = 2$.

Exercise 3.23
SAT scores closely follow the normal model with mean $\mu = 1500$ and standard deviation $\sigma = 300$. (a) About what percent of test takers score 900 to 2100? (b) What percent score between 1500 and 2100?

3.2 Evaluating the normal approximation

Many processes can be well approximated by the normal distribution. We have already seen two good examples: SAT scores and the heights of US adult males. While using a normal model can be extremely convenient and helpful, it is important to remember normality is

First draw the pictures. To find the area between $Z = 1$ and $Z = 1$, use the normal probability table to determine the areas below $Z = 1$ and above $Z = 1$. Next, verify the area between $Z = 1$ and $Z = 1$ is about 0.68. Repeat this for $Z = 2$ to $Z = 2$.

(a) 900 and 2100 represent two standard deviations above and below the mean, which means about 95% of test takers will score between 900 and 2100. (b) Since the normal model is symmetric, then half of the test takers from part (a) ($\frac{95\%}{2} = 47.5\%$ of all test takers) will score 900 to 1500 while 47.5% score between 1500 and 2100.

Source: Diez et al. 2011
4.2. CONFIDENCE INTERVALS

The area between \(-z^*\) and \(z^*\) increases as \(|z^*|\) becomes larger. If the confidence level is 99%, we choose \(z^*\) such that 99% of the normal curve is between \(-z^*\) and \(z^*\), which corresponds to 0.5% in the lower tail and 0.5% in the upper tail: \(z^* = 2.58\).
Confidence intervals

If the sampling distribution is approximately normal, fractions of the point estimates are contained within $Z$-scores:

- For a 95% CI: $\bar{X} - 1.96 \cdot SEM, \bar{X} + 1.96 \cdot SEM$
- For a 99% CI: $\bar{X} - 2.58 \cdot SEM, \bar{X} + 2.58 \cdot SEM$

Wider intervals trade precision for additional confidence.

Margin of error

The margin of error of the interval $\bar{X} \pm Z \cdot SEM$ is $Z \cdot SEM$.

Sanity check

Confidence intervals are estimations of the population parameter; they say nothing of the sample itself.
Homework

Read CK-12 handbook ch. 7–8 for next week

and enjoy the rest of your day.

Note: final stats exam will cover confidence intervals (Ch. 7) and hypothesis tests (Ch. 8). Histograms are part of the topic.