1 Sample Size

- In some cases, experiemtners would like the results of an experiment to be accurate within a certain amount, say $b > 0$.

- In the case of large sample experiments, this is an achievable request
- With large samples we know that $\overline{X}$ will be our estimate for $E[X_i]$ and that our $1 - \alpha$ CI will be
$$(\overline{X} - \frac{\sigma}{\sqrt{n}} Z_{\alpha/2}, \overline{X} + \frac{\sigma}{\sqrt{n}} Z_{\alpha/2})$$
- We can think of the $\frac{\sigma}{\sqrt{n}} Z_\alpha$ part of this CI as the measurement of accuracy of our $\overline{X}$ estimate
- So, if we want our estimate to be within b units of measurement of the true value of the mean (with $1 - \alpha$ probability), then that means that we want
$$\begin{aligned} \frac{\sigma}{\sqrt{n}} Z_{\alpha/2} &= b \\ \Rightarrow n &= \frac{\sigma^2}{b^2} Z^2_{\alpha/2} \end{aligned}$$

  For fixed $\alpha, b, \text{and } \sigma$
- We don't usually have σ so we often estimate it with previous estimates, or based on informed beliefs

1.1 Example

Suppose that you want to estimate the average daily yield of a chemical process μ. If we assume that the standard deviation of this process is about 20 tons for a given day, then how big of a sample would we want to take in order to have our estimate be within 5 tons of the true mean μ with a 95% probability?
Solution:
Here, $\alpha = .05 \Rightarrow Z_{\alpha/2} \approx 1.96$. We also have $b = 5$. So we would want our sample size to be approximately

$$\begin{aligned} n &= \frac{\sigma^2}{b^2} Z^2_{\alpha/2} \\ &\approx \frac{400}{25} 1.96^2 \\ &\approx 61.4656 \end{aligned}$$

So, we would want ot take a sample of size $n = 62$ to create the desired precision

1.2 Exercise

Suppose You are a chemist who is trying to synthesize a specific compund. You have been working with a new technique and you think that this process can turn 60% of the imput compunds into the desired synthesized compund, and want to attempt the process enough times to get an estimate that is within .04 of the true proportion that is converted. how many times should you execute the process to get the desired precision?

1.3 Solution

Suppose You are a chemist who is trying to synthesize a specific compund. You have been working with a new technique and you think that this process can turn 60% of the imput compunds into the desired synthesized compund, and want to attempt the process enough times to get an estimate that is within .04 of the true proportion that is converted. how many times should you execute the process to get the desired precision?
Solution:
Because this is a proportion problem, we can estimate $V[X_i] = p(1-p)$ with $.6(1-.6)$ since we think that p is around .6. This gives us:

$$\begin{aligned} n &= \frac{\sigma^2}{b^2} Z^2_{\alpha/2} \\ &\approx \frac{.6(.4)}{(.04)^2} 1.96^2 \\ &\approx 576.24 \end{aligned}$$

So, we would want to run the process 577 times to get the desired precision.

2 Normal Population CIs

Now we will focus on one of the most prolific CI problems in statistics, the small sample CI in normal population situations.

2.1 Examples

Let $X_1, ..., X_n$ be i.i.d. $N(\mu_x, \sigma_x^2)$, $Y_1, ..., Y_m$ be i.i.d. $N(\mu_y, \sigma_y^2)$, and let all of the Xs and Ys be independent.
Assume that σ_x^2 and σ_y^2 are both known

1. Derive a two sided $1-\alpha$ CI for μ_x based on $X_1, ..., X_n$
   Solution:
   We know that

$$\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n})$$

$$\begin{aligned}
\Rightarrow \frac{\overline{X}-\mu_x}{\sqrt{\sigma_x^2/n}} &\sim N(0,1) \\
\Rightarrow P(-Z_{\alpha/2} < \frac{\overline{X}-\mu_x}{\sqrt{\sigma_x^2/n}} < Z_{\alpha/2}) &= 1-\alpha \\
\Rightarrow P(\overline{X}-\sqrt{\sigma_x^2/n}Z_{\alpha/2} < \mu_x < \overline{X}+\sqrt{\sigma_x^2/n}Z_{\alpha/2}) &= 1-\alpha
\end{aligned}$$

So, our $1-\alpha$ CI is $(\overline{X}-\sqrt{\sigma_x^2/n}Z_{\alpha/2}, \overline{X}+\sqrt{\sigma_x^2/n}Z_{\alpha/2})$

2. Derive a two sided $1-\alpha$ CI for $\mu_x-\mu_y$ based on $X_1, ..., X_n, Y_1, ..., Y_m$
Solution:
We know that

$$\begin{aligned}
\overline{X} \sim N(\mu_x, \frac{\sigma_x^2}{n}), \qquad \overline{Y} &\sim N(\mu_y, \frac{\sigma_y^2}{m}) \\
\Rightarrow \overline{X}-\overline{Y} &\sim N(\mu_x-\mu_y, \frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}) \\
\Rightarrow \frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}} &\sim N(0,1)
\end{aligned}$$

$$\begin{aligned}
\Rightarrow P(-Z_{\alpha/2} < \frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}} < Z_{\alpha/2}) &= 1-\alpha \\
\Rightarrow P(\overline{X}-\overline{Y}-\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}Z_{\alpha/2} < \mu_x-\mu_y < \overline{X}-\overline{Y}+\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}Z_{\alpha/2}) &= 1-\alpha
\end{aligned}$$

So, our $1-\alpha$ CI is $(\overline{X}-\overline{Y}-\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}Z_{\alpha/2}, \overline{X}-\overline{Y}+\sqrt{\frac{\sigma_x^2}{n}+\frac{\sigma_y^2}{m}}Z_{\alpha/2})$

2.2 Exercises

Let $X_1, ..., X_n$ be i.i.d. $N(\mu_x, \sigma_x^2)$, $Y_1, ..., Y_m$ be i.i.d. $N(\mu_y, \sigma_y^2)$, and let all of the Xs and Ys be independent.
Assume that $\sigma_x = \sigma_y = \sigma$ is unkown.

1. Construct a two sided $1-\alpha$ CI for μ_x based on $X_1, ..., X_n$
   a) First contruct a pivot. We know that $\overline{X} \sim N(\mu_x, \sigma^2)$, $\frac{n-1}{\sigma^2}S_x^2 \sim \chi^2_{n-1}$ and that $\overline{X}$ and S_x^2 are independent. Using this information, try to construct a statistic that is a function of μ_x (but can be shown to not depend upon σ that has a known distirbution that does not depend on any unkown values.

   b) Using this pivot construct the CI

2. Construct a two sided $1-\alpha$ CI for $\mu_x - \mu_y$ based on $X_1, ..., X_n$

   a) First, we need a pivot. We know that $\overline{X} - \overline{Y} \sim N(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m})$. We also know that $\frac{n-1}{\sigma^2} S_x^2 \sim \chi^2_{n-1}$ and $\frac{m-1}{\sigma^2} S_y^2 \sim \chi^2_{m-1}$. Find the distirbution of $\frac{n-1}{\sigma^2} S_x^2 + \frac{m-1}{\sigma^2} S_y^2$.

   b) Let $S_p^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2 + \sum_{i=1}^m (Y_i - \overline{Y})^2}{n+m-2}$ Show that $\frac{n+m-2}{\sigma^2} S_p^2 = \frac{n-1}{\sigma^2} S_x^2 + \frac{m-1}{\sigma^2} S_y^2$

   c) It can be shown that $\frac{n+m-2}{\sigma^2} S_p^2 = \frac{n-1}{\sigma^2} S_x^2 + \frac{m-1}{\sigma^2} S_y^2$ and $\overline{X} - \overline{Y}$ are independent. Knowing the distribution of $\frac{n-1}{\sigma^2} S_x^2 + \frac{m-1}{\sigma^2} S_y^2 = \frac{n+m-2}{\sigma^2} S_p^2$, construct a pivot for $\mu_x - \mu_y$.

   d) Construct the $1-\alpha$ CI for $\mu_x - \mu_y$

2.3 Solutions

Let $X_1, ..., X_n$ be i.i.d. $N(\mu_x, \sigma_x^2)$, $Y_1, ..., Y_m$ be i.i.d. $N(\mu_y, \sigma_y^2)$, and let all of the Xs and Ys be independent.
Assume that $\sigma_x = \sigma_y = \sigma$ is unkown.

1. Construct a two sided $1-\alpha$ CI for μ_x based on $X_1, ..., X_n$

   a) First contruct a pivot. We know that $\overline{X} \sim N(\mu_x, \sigma^2)$, $\frac{n-1}{\sigma^2} S_x^2 \sim \chi^2_{n-1}$ and that $\overline{X}$ and S_x^2 are independent. Using this information, try to construct a statistic that is a function of μ_x (but can be shown to not depend upon σ that has a known distirbution that does not depend on any unkown values.
   Solution:

$$\begin{aligned}
\overline{X} &\sim N(\mu_x, \frac{\sigma^2}{n}) \\
\Rightarrow \sqrt{n}\frac{\overline{X} - \mu_x}{\sigma} = \frac{\overline{X} - \mu_x}{\sqrt{\sigma^2/n}} &\sim N(0,1) \\
\rightarrow \frac{n-1}{\sigma^2} S_x^2 &\sim \chi^2_{n-1} \\
\Rightarrow \frac{\sqrt{n}(\overline{X} - \mu_x)/\sigma}{\sqrt{\frac{n-1}{\sigma^2} S_x^2/(n-1)}} &\sim t_{n-1} \\
\Rightarrow \frac{(\overline{X} - \mu_x)}{\sqrt{S_x^2/n}} &\sim t_{n-1}
\end{aligned}$$

   b) Using this pivot construct the CI. *Hint:* The t distribution is symmetric about 0 for all degreees of freedom.

Solution:
Let $-t_{n-1,\alpha/2}$ and $t_{n-1,\alpha/2}$ be the $\alpha/2$ and $1-\alpha/2$ percentiles of t_{n-1}.

$$\begin{aligned}
\frac{(\overline{X}-\mu_x)}{\sqrt{S_x^2/n}} &\sim t_{n-1} \\
\Rightarrow P(-t_{n-1,\alpha/2} < \frac{(\overline{X}-\mu_x)}{\sqrt{S_x^2/n}} < t_{n-1,\alpha/2}) &= 1-\alpha \\
\Rightarrow P(\overline{X}-\sqrt{S_x^2/n}t_{n-1,\alpha/2} < \mu_x < \overline{X}+\sqrt{S_x^2/n}t_{n-1,\alpha/2}) &= 1-\alpha
\end{aligned}$$

So, our $1-\alpha$ CI is $(\overline{X}-\sqrt{S_x^2/n}t_{n-1,\alpha/2}, \overline{X}+\sqrt{S_x^2/n}t_{n-1,\alpha/2})$

2. Construct a two sided $1-\alpha$ CI for $\mu_x-\mu_y$ based on $X_1, ..., X_n$

a) First, we need a pivot. We know that $\overline{X}-\overline{Y} \sim N(\mu_x-\mu_y, \frac{\sigma^2}{n}+\frac{\sigma^2}{m})$. We also know that $\frac{n-1}{\sigma^2}S_x^2 \sim \chi^2_{n-1}$ and $\frac{m-1}{\sigma^2}S_y^2 \sim \chi^2_{m-1}$. Find the distirbution of $\frac{n-1}{\sigma^2}S_x^2+\frac{m-1}{\sigma^2}S_y^2$.
Solution:

$$\begin{aligned}
\frac{n-1}{\sigma^2}S_x^2 \sim \chi^2_{n-1} &\equiv \Gamma(\frac{n-1}{2},2) \\
\frac{m-1}{\sigma^2}S_y^2 \sim \chi^2_{m-1} &\equiv \Gamma(\frac{m-1}{2},2) \\
\Rightarrow \frac{n-1}{\sigma^2}S_x^2+\frac{m-1}{\sigma^2}S_y^2 \sim \Gamma(\frac{n+m-2}{2},2) &\equiv \chi^2_{n+m-2}
\end{aligned}$$

Since all of the Xs and Ys are independent $\uparrow$

b) Let $S_p^2 = \frac{\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{i=1}^m (Y_i-\overline{Y})^2}{n+m-2}$ Show that $\frac{n+m-2}{\sigma^2}S_p^2 = \frac{n-1}{\sigma^2}S_x^2+\frac{m-1}{\sigma^2}S_y^2$
Solution:

$$\begin{aligned}
\frac{n-1}{\sigma^2}S_x^2+\frac{m-1}{\sigma^2}S_y^2 &= \frac{n-1}{\sigma^2}\frac{\sum_{i=1}^n (X_i-\overline{X})^2}{n-1}+\frac{m-1}{\sigma^2}\frac{\sum_{i=1}^m (Y_i-\overline{Y})^2}{m-1} \\
&= \frac{\sum_{i=1}^n (X_i-\overline{X})^2}{\sigma^2}+\frac{\sum_{i=1}^m (Y_i-\overline{Y})^2}{\sigma^2} \\
&= \frac{\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{i=1}^n (Y_i-\overline{Y})^2}{\sigma^2} \\
&= \frac{n+m-2}{\sigma^2}\frac{\sum_{i=1}^n (X_i-\overline{X})^2+\sum_{i=1}^m (Y_i-\overline{Y})^2}{n+m-2} \\
&= \frac{n+m-2}{\sigma^2}S_p^2
\end{aligned}$$

c) It can be shown that $\frac{n+m-2}{\sigma^2}S_p^2 = \frac{n-1}{\sigma^2}S_x^2 + \frac{m-1}{\sigma^2}S_y^2$ and $\overline{X} - \overline{Y}$ are independent. Knowing the distribution of $\frac{n-1}{\sigma^2}S_x^2 + \frac{m-1}{\sigma^2}S_y^2 = \frac{n+m-2}{\sigma^2}S_p^2$, construct a pivot for $\mu_x - \mu_y$.
Solution:

$$\begin{aligned}
\overline{X} - \overline{Y} &\sim N(\mu_x - \mu_y, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}) \\
\Rightarrow \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{m}}} &\sim N(0,1) \\
\rightarrow \frac{n+m-2}{\sigma^2}S_p^2 &\sim \chi^2_{n+m-2}
\end{aligned}$$

$\rightarrow \overline{X} - \overline{Y}$ and $\frac{n+m-2}{\sigma^2}S_p^2$ are independent

$$\begin{aligned}
\Rightarrow \frac{\frac{\overline{X}-\overline{Y}-(\mu_x-\mu_y)}{\sqrt{\frac{\sigma^2}{n}+\frac{\sigma^2}{m}}}}{\sqrt{\frac{n+m-2}{\sigma^2}S_p^2/(m+n-2)}} &\sim t_{m+n-2} \\
\Rightarrow \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} &\sim t_{n+m-2}
\end{aligned}$$

d) Construct the $1 - \alpha$ CI for $\mu_x - \mu_y$

$$\frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} \sim t_{n+m-2}$$

$$\Rightarrow P(-t_{n+m-2,\alpha/2} < \frac{\overline{X} - \overline{Y} - (\mu_x - \mu_y)}{\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}} < t_{n+m-2,\alpha/2}) = 1 - \alpha$$

$$\Rightarrow P(-t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})} < \overline{X}-\overline{Y}-(\mu_x-\mu_y) < t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}) = 1-\alpha$$

$$\Rightarrow P(\overline{X}-\overline{Y}-t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})} < \mu_x-\mu_y < \overline{X}-\overline{Y}+t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}) = 1-\alpha$$

So, our $1-\alpha$ CI is $(\overline{X}-\overline{Y}-t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})}, \overline{X}-\overline{Y}+t_{n+m-2,\alpha/2}\sqrt{S_p^2(\frac{1}{n} + \frac{1}{m})})$