Confidence intervals
Prevalence (with 95% CI bars) of obesity among New York City public elementary school children, by sex and race/ethnicity, 2003. (source: CDC.GOV)

What do those bars actually mean?
Patterns of somatic mutation in human cancer genomes

The numbers of passenger and driver mutations present can be estimated from these results (see Supplementary Methods). Of the 921 base substitutions in the primary screen, 763 (95% confidence interval, 675–858) are estimated to be passenger mutations. Therefore, the large majority of mutations found through sequencing cancer genomes are not implicated in cancer development, even when the search has been targeted to the coding regions of a gene family of high candidature. However, there are an estimated 158 driver mutations (95% confidence interval, 63–246), accounting for the observed positive selection pressure. These are estimated to be distributed in 119 genes (95% confidence interval, 52–149). The number of samples containing a driver mutation is estimated to be 66 (95% confidence interval, 36–77). The results, therefore, provide statistical evidence for a large set of mutated protein kinase genes implicated in the development of about one-third of the cancers studied.
Two-sided confidence intervals

• Calculated based on the sample $X_1, X_2, ..., X_n$
• Characterized by:
  – lower- and upper- confidence limits $L$ and $R$
  – the confidence coefficient $1-\alpha$
• Objective: for two-sided confidence interval, find $L$ and $R$ such that
  – $\text{Prob}(\mu > R) = \alpha/2$
  – $\text{Prob}(\mu < L) = \alpha/2$
  – Therefore, $\text{Prob}(L < \mu < R) = 1-\alpha$
• For one-sided confidence interval, say, lower bound of $\mu$, find $R$ that
  – $\text{Prob}(\mu > R) = \alpha$
• **Assume standard deviation sigma is known**
Consider $1 - \alpha = 95\% = 0.95$

$\alpha = 0.05$; $\frac{\alpha}{2} = 0.025$

$z_{\frac{\alpha}{2}} = 1.96 \Rightarrow Prob(T > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

$Prob(-\frac{z_{\frac{\alpha}{2}}}{\sqrt{n}} < \frac{X - \mu}{\sqrt{n}} < \frac{z_{\frac{\alpha}{2}}}{\sqrt{n}}) = 1 - \alpha$

$Prob\left(\frac{\bar{X} - 2z_{\frac{\alpha}{2}}}{\sqrt{n}} < \mu < \bar{X} + \frac{2z_{\frac{\alpha}{2}}}{\sqrt{n}}\right) = 1 - \alpha$

For one-sided lower bound on $\mu$

$Prob\left(\frac{\bar{X} - \mu}{\sqrt{n}} < z_{\alpha}\right) \rightarrow$

$\mu > \bar{X} - z_{\alpha} \frac{2}{\sqrt{n}}$

$z_{\alpha} = 1.65 < z_{0.025} = 1.96$
Exercise

Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface.

Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm².

Assume that the standard deviation is known to be 0.66 dyne-cm²

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm² what is the minimal number of samples should be?
Ishikawa et al. (Journal of Bioscience and Bioengineering 2012) studied the force with which bacterial biofilms adhere to a solid surface. Five measurements for a bacterial strain of Acinetobacter gave readings 2.69, 5.76, 2.67, 1.62, and 4.12 dyne-cm\(^2\). Assume that the standard deviation is known to be 0.66 dyne-cm\(^2\)

(a) Find 95% confidence interval for the mean adhesion force

(b) If scientists want the width of the confidence interval to be below 0.55 dyne-cm\(^2\) what the minimal number of samples should be?

\[ n = 5 \quad \sigma = 0.66 \quad \bar{x} = 3.372, \quad z = 1.96 \]

\[ \bar{x} - z\sigma / \sqrt{n} \leq \mu \leq \bar{x} + z\sigma / \sqrt{n} \]

\[ 3.372 - 1.96(0.66 / \sqrt{5}) \leq \mu \leq 3.372 + 1.96(0.66 / \sqrt{5}) \]

\[ 2.79 \leq \mu \leq 3.95 \]

b) Width is \[ 2z\sigma / \sqrt{n} = 0.55 \], therefore \[ n = [2z\sigma / 0.55]^2 = [2(1.96)(0.66) / 0.55]^2 = 22.13 \]
Round up to \[ n = 23 \].
2. (6 points) The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a **90% confidence interval** for the population mean of the assembly time.

What Z should I look for in the table?

A. $\Phi(Z)=0.9$
B. $\Phi(Z)=0.05$
C. $\Phi(Z)=0.95$
D. $\Phi(Z)=0.1$
E. I have no idea

Get your i-clickers
2. (6 points) The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a 90% confidence interval for the population mean of the assembly time.

What Z should I look for in the table?

A. \( \Phi(Z) = 0.9 \)
B. \( \Phi(Z) = 0.05 \)
C. \( \Phi(Z) = 0.95 \)
D. \( \Phi(Z) = 0.1 \)
E. I have no idea

Get your i-clickers
2. (6 points) The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a 90% confidence interval for the population mean of the assembly time.
2. **(6 points)** The operations manager of a large production plant would like to estimate the mean amount of time a worker takes to assemble a new electronic component. Assume that the standard deviation of this assembly time is 3.6 minutes. After observing a sample of 100 workers assembling similar devices, the manager noticed that their average time was 16.2 minutes. Construct a **90% confidence interval** for the population mean of the assembly time.

**Answer:** Let \( \mu \) denote the mean assembly time (in minutes). We want a 90% confidence interval for \( \mu \) based on the following information: \( n = 100, \bar{X} = 16.2, \alpha = 0.1, \sigma = 3.6 \). Since \( \sigma \) is known, we can use normal distribution to calculate confidence interval:

\[
\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = 16.2 \pm (1.65) \frac{3.6}{\sqrt{10}} = [15.61, 16.79]
\]
8-2 Confidence Interval on the Mean of a Normal Distribution, Variance Known

Figure 8-1  Repeated construction of a confidence interval for $\mu$.

Figure 8-1  Repeated construction of a confidence interval for $\mu$. 1 out of 16 consistent with 5% or 1 out of 20.
Matlab exercise

- 1000 labs measured average P53 gene expression.
- Each lab used a sample with n=20 drawn from the Gaussian distribution with mu=3; sigma=2;
- Each lab found 95% confidence estimates of the population mean mu based on its own sample only
- Count the number of labs, where the true population mean (mu=3) lies outside their 95% confidence bounds
- You should get ~50 labs out of 1000 labs
How I did it

- \( n=20; \ k_{\text{labs}}=1000; \)
- \( \text{rand\_table}=2.*\text{randn}(n, k_{\text{labs}})+3; \)
- \( \text{sample\_mean} = \text{mean}(\text{rand\_table}, 1); \)
- \( \text{CI\_low} = \text{sample\_mean} - 1.96.*2./\sqrt{n}; \)
- \( \text{CI\_high} = \text{sample\_mean} + 1.96.*2./\sqrt{n}; \)
- \( \text{k\_above} = \text{sum}(3>\text{CI\_high}) \)
- \( \text{k\_below} = \text{sum}(3<\text{CI\_low}) \)
- \( \text{figure}; \ \text{ndisp}=100; \ \text{errorbar}(1: \text{ndisp}, \ \text{sample\_mean}(1: \text{ndisp}), \ \text{ones}(\text{ndisp},1).*1.96.*2./\sqrt{n},'ko'); \)
- \( \text{hold on}; \ \text{plot}(1: \text{ndisp}, 3.*\text{ones}(\text{ndisp},1),'r-'); \)
Is P53 gene expressed at a lower level in cancer patients than in healthy people?

- We are interested if a P53 gene expression is lowered in population of cancer patients compared to the healthy population.
- We know that mean gene expression in the healthy population is $\mu_h = 50$ mRNAs/cell. We are interested in deciding whether or not the mean expression in cancer population is lower than in healthy population. Called hypothesis $H_1$. Here $H_1$ is one-sided. If we asked: cancer not equal healthy – two-sided hypothesis.
- Assume we have a sample of 100 cancer patients with sample mean $\bar{X} = 48$ mRNAs/cell and sample standard deviation $S = 10$ mRNA/cell.
- Can we use our sample to reject the "strawman" or null hypothesis $H_0$: cancer = healthy?
### Two types of errors

<table>
<thead>
<tr>
<th></th>
<th>decide $H_0$</th>
<th>decide $H_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>true $H_0$</strong></td>
<td>Correct action $1 - \alpha$</td>
<td>Type I error $\alpha$</td>
</tr>
<tr>
<td>probability</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>true $H_1$</strong></td>
<td>Type II error $\beta$</td>
<td>Correct action power $= 1 - \beta$</td>
</tr>
<tr>
<td>probability</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
problem above, we can say that $H_0: \mu = 50$ was rejected at the 0.05 level of significance. This statement of conclusions is often inadequate because it gives the decision maker no idea about whether the computed value of the test statistic was just barely in the rejection region or whether it was very far into this region. Furthermore, stating the results this way imposes the predefined level of significance on other users of the information. This approach may be unsatisfactory because some decision makers might be uncomfortable with the risks implied by $\alpha = 0.05$.

The **P-value** is the smallest level of significance that would lead to rejection of the null hypothesis $H_0$ with the given data.

It is customary to call the test statistic (and the data) significant when the null hypothesis $H_0$ is rejected; therefore, we may think of the P-value as the smallest level $\alpha$ at which the data are significant. Once the P-value is known, the decision maker can determine how significant the data are without the data analyst formally imposing a preselected level of significance.
\[ \mu_h = 50 \]

\[ H_0 : \mu_c = \mu_h \]

\[ n = 100, \ \bar{X} = 48, \ \sigma = 10 \]

One-sided hypothesis \( H_1 : \mu_c < \mu_h \)

\[ \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \]

\[ \bar{X} = 48 \]

\[ P-value(\bar{X} = 48 \mid H_0 : \mu_c = \mu_h) = \text{Prob}(\bar{X} \leq 48) \approx 2.5\% \approx 0.025 \]

If \( H_1 : \mu_c \neq \mu_h \):

\[ P-value = \text{Prob}(\bar{X} \leq 48) + \text{Prob}(\bar{X} \geq 52) = 0.05 \]

Two-sided hypothesis
\( \mu_h = 50 \)

\( H_0: \mu_c = \mu_h \)

\( n = 100, \ X = 48, \ S = 10 \)

\( H_1: \mu_c < \mu_h \)

\[ \sigma_X = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \]
\[ H_0: \mu_c = \mu_h \]

\[ H_1: \mu_c < \mu_h \]

\[ n=100, \; \bar{X}=48, \; S=10 \]

\[ \sigma_x = \frac{S}{\sqrt{n}} = \frac{10}{\sqrt{100}} = 1 \]

Set p-value threshold:

\[ \alpha = 2.5\% \]

Type II error

\[ \beta = P(\text{Accept } H_0 \mid H_1 \text{ is true}) = \int_{-\infty}^{\bar{X}+1.96} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x-48.04)^2}{2}\right) \, dx = 1 - \Phi(1.04) = 15\% \]

\[ \alpha = 1 - \Phi(1.96) = 2.5\% \]
Generalizations

• What if $H_1$ is a two-sided hypothesis?

• A: P-value is $2(1-\Phi(|Z|))$, where $Z=(\bar{X}-\mu_0)/[S/\sqrt{n}]$

  Compare to: For one sized $\mu_1 > \mu_0$ it is $1-\Phi(Z)$
  For one sized $\mu_1 < \mu_0$ it is $\Phi(Z)$

• If $\alpha$ is given use $\mu_0 +/- z_{\alpha/2}$ as thresholds to reject the null hypothesis

• What if the sample size $n$ is small (say $n<10$):

• A: Use t-distribution with n-1 degrees of freedom for 2-sided $P$-value=$2(1-CDF_{Tdist}(|T|))$

  where $T=(\bar{X}-\mu_0)/[S/\sqrt{n}]$.

• For given $\alpha$ use $\mu_0 +/- t_{\alpha/2,n-1}$ to reject the null hypothesis
Type II error for two-sided hypothesis

\[ \text{Prob Type II error} = \beta \]

for small \( n \):

\[ M_0 - t_{\alpha/2, n-1} \quad M_0 - t_{\beta/2, n-1} \quad M_0 \quad M_0 + t_{\alpha/2, n-1} \quad M_0 + t_{\beta/2, n-1} \]
Type II Error and Choice of Sample Size

Assume you know the minimum $\delta = |\mu_1 - \mu_0|$ that you care about. What is the minimal sample you should use to separate $H_0$ and $H_1$ hypotheses if your tolerance to type I and type II errors is $\alpha$ and $\beta$?

\[
\frac{\delta \sqrt{n}}{\sigma} = z_{\alpha/2} + z_\beta
\]

\[
n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2} \quad \text{where} \quad \delta = \mu - \mu_0 \quad (9-22)
\]