2019-2

Introduction to Biostatistics - Lecture 2: Statistical Inference Procedures

Jonggyu Baek
University of Massachusetts Medical School

Follow this and additional works at: https://escholarship.umassmed.edu/liberia_peer

Part of the Biostatistics Commons, Family Medicine Commons, Infectious Disease Commons, Medical Education Commons, and the Public Health Commons

Repository Citation
https://escholarship.umassmed.edu/liberia_peer/9

This material is brought to you by eScholarship@UMMS. It has been accepted for inclusion in PEER Liberia Project by an authorized administrator of eScholarship@UMMS. For more information, please contact Lisa.Palmer@umassmed.edu.
Introduction to Biostatistics

2/28/2019

Jonggyu Baek, PhD
Lecture 2:

• Statistical Inference Procedures
  – Hypothesis test for population average
  – Hypothesis test for comparing means
  – Power and sample size
Statistical Inference

Two broad areas of statistical inference:

• **Estimation:** Use sample statistics to estimate the unknown population parameter.
  
  – **Point Estimate:** the best single value to describe the unknown parameter.
  
  – **Standard Error (SE):** standard deviation of the sample statistic. Indicates how precise is the point estimate.
  
  – **Confidence Interval (CI):** the range with the most probable values for the unknown parameter with a \((1-\alpha)\%\) level of confidence.

• **Hypothesis Testing:** Test a specific statement (assumption) about the unknown parameter.
Statistical Inference for population average $\mu$

Estimation: Point Estimate & Standard Error

• Suppose $X$ a variable (e.g., systolic BP, hypertension, # of prior complications) from a population of size $N$ with average $\mu$ and standard deviation $\sigma$.

• We select a random sample $x_1, x_2, \ldots, x_n$ of size $n$

• **Point Estimate** of $\mu$ : $\bar{x}$

• **Standard error** of $\bar{x}$ : Standard Deviation of all possible $\bar{x}$ ‘s

• From the central limit theorem (CLT), for $n$ large ($n \geq 30$):

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

• If $\sigma$ also unknown we can estimate from the sample standard deviation $s$. 
The Central Limit Theorem (CLT)

Suppose X from a population (N) with $\mu$ and $\sigma$.

- If we take random samples ($n$) with replacement from the population, for large “$n$” the distribution of the sample mean $\bar{x}$ is approximately normally distributed with $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$, i.e.:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Importance:
- The distribution of the sample mean ($\bar{x}$) is approximately normal even if X does not follow $N(\mu, \sigma)$.
- Sample mean is very useful for statistical inference.
Normal Distribution

Examples:

- $N(0, 1)$
- $N(2, 1)$
- $N(0, 2)$
- $N(2, 2)$
The Standard Normal Distribution

$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$ can be transformed to a $Z \sim N(0, 1)$:

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

- N(0, 1) is called the standard normal distribution
- Z is the standardized value of $\bar{x}$
- Standardized values make comparable variables that are measured in different units, or have different variability
Statistical Inference for population average $\mu$

**Estimation: Confidence Interval**

- **Confidence Interval (CI):** a range of values that are likely to cover the true parameter value with a level of confidence $(1-\alpha)\%$ assigned to it. The most common choice for $\alpha$ is 5%.

- Usually CIs are symmetric around the point estimate.

- From the central limit theorem (CLT), for $n$ large ($n \geq 30$):

  $$\overline{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- Hence,

  $$Z = \frac{\overline{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$
Statistical Inference for population average $\mu$

**Estimation: Confidence Interval**

- E.g., $(1-\alpha)=95\%$ CI for $\mu$

95% CI for average $\mu$: $[\bar{x} - 1.96 \cdot (\sigma / \sqrt{n}), \bar{x} + 1.96 \cdot (\sigma / \sqrt{n})]$

How we derived its 95% CI?

- 95% of Z around 0 is between -1.96 and 1.96
  [or $Z_{0.025} = -1.96$ and $Z_{0.975} = 1.96$]

- Remember that Z does not have any scale because it is standardized. We need the scale back to calculate 95% CI.
Statistical Inference for population average $\mu$

**Estimation: Confidence Interval**

- Based on the percentiles of the N(0,1) there are some commonly reported CIs:

<table>
<thead>
<tr>
<th>(1-(\alpha))% CI</th>
<th>(\alpha)</th>
<th>(\alpha/2)</th>
<th>1-(\alpha/2)</th>
<th>(Z_{\alpha/2})</th>
<th>(Z_{1-\alpha/2})</th>
</tr>
</thead>
<tbody>
<tr>
<td>80%</td>
<td>20</td>
<td>10</td>
<td>90</td>
<td>-1.28</td>
<td>1.28</td>
</tr>
<tr>
<td>90%</td>
<td>10</td>
<td>5</td>
<td>95</td>
<td>-1.64</td>
<td>1.64</td>
</tr>
<tr>
<td>95%</td>
<td>5</td>
<td>2.5</td>
<td>97.5</td>
<td>-1.96</td>
<td>1.96</td>
</tr>
<tr>
<td>99%</td>
<td>1</td>
<td>0.5</td>
<td>99.5</td>
<td>-2.58</td>
<td>2.58</td>
</tr>
</tbody>
</table>
Example of CIs: The Framingham Heart Study

- Can you calculate 95% CIs based only on descriptive statistics for the systolic blood pressure?

```
library(psych)
describe(dat1$sysbp)
```

```
> library(psych)
> describe(dat1$sysbp)

   vars n mean  sd median trimmed   mad min max range skew kurtosis   se
     x1   1 11627 136.32 22.8   132 134.34 20.76 83.5 295 211.5 0.94 1.37 0.21
```

95% CI : \[ \bar{x} - 1.96\cdot(\sigma/\sqrt{n}) , \bar{x} + 1.96\cdot(\sigma/\sqrt{n}) \]

= [136.32- 1.96\cdot0.21 , 136.32+ 1.96\cdot0.21]

= [135.91, 136.73]
Example of CIs: The Framingham Heart Study

- Is there any way to calculate 95% CI directly?

```r
t.test(dat1$sysbp)

> t.test(dat1$sysbp)

One sample t-test

data:  dat1$sysbp
t = 644.76, df = 11626, p-value < 2.2e-16
classical hypothesis: true mean is not equal to 0

95 percent confidence interval:
135.9097 136.7386
sample estimates:
mean of x
136.3241
```
Hypothesis Testing for the mean $\mu$

• Suppose $X$ continuous from a population with mean $\mu$ and standard deviation $\sigma$.

• **What is the value of $\mu$?**

• We select a random sample from that population and try to make inference about $\mu$. 
Key Concepts in Hypothesis Testing

- **Null hypothesis (H₀):**
  - An explicit statement about an unknown parameter the validity of which you wish to test, e.g., \( \mu = \mu_0 \)

- **Alternative hypothesis (H₁):**
  - An alternative statement about the unknown parameter used to compare your null with, e.g.,
    - \( \mu \neq \mu_0 \) (two-sided test)
    - \( \mu < \mu_0 \) (one-sided test)
    - \( \mu > \mu_0 \) (one-sided test)

- **Errors:**
  - Type I: reject \( H_0 \) | \( H_0 \) is true (crucial)
  - Type II: do not reject \( H_0 \) | \( H_1 \) is true (moderate)
Think of **Type I** error as the "**presumption of innocence**" according to which "everyone is presumed innocent until proven guilty":

“It is better that ten guilty persons escape than that one innocent suffer” from the principle of Blackstone formula:

- $H_0$: a person is innocent
- $H_1$: a person is guilty

• Without enough evidences, a person is innocent

What about this?

- $H_0$: a person is guilty
- $H_1$: a person is innocent

• Without enough evidences, a person is guilty
Hypothesis Testing for the mean $\mu$

- **What is the value of $\mu$?** (e.g., the population mean of systolic BP is 136.)

- **Hypothesis Test:**
  
  $H_0$: $\mu = \mu_0 (=136)$
Hypothesis Testing for the mean $\mu$

- What is the value of $\mu$?

- Hypothesis Test:
  $H_0: \mu = \mu_0 (=136)$

- Random sample:
  $\bar{x}$
Hypothesis Testing for the mean $\mu$

- What is the value of $\mu$?

- Hypothesis Test:
  $H_0: \mu = \mu_0$ (?)

- Random sample:
  $\bar{x}$

- If $\bar{x}$ close to $\mu_0 \rightarrow H_0$ probable
- If $\bar{x}$ far from $\mu_0 \rightarrow H_0$ not probable
Hypothesis Testing for the mean $\mu$

• What is the value of $\mu$?

• Hypothesis Test:
  $H_0: \mu = \mu_0$ (?)

• Random sample:
  $\bar{x}$

• If $\bar{x}$ close to $\mu_0$ $\rightarrow$ $H_0$ probable
• If $\bar{x}$ far from $\mu_0$ $\rightarrow$ $H_0$ not probable
Key Concepts in Hypothesis Testing

• **Test Statistic:**
  – A summary measure of your sample, with known distribution under $H_0$, used for testing the null hypothesis ($H_0$), e.g.,
  \[
  Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)
  \]

• **Critical points:**
  – Values (percentiles) of the known distribution of the test statistic above or below which the probability of Type I Error is $\alpha\%$, e.g.,
  \[Z_\alpha, \ Z_{\alpha/2}, \ Z_{1-\alpha/2}, \ t_{1-\alpha/2, \text{d.f.}}, \ etc.\]
Statistical Inference for population average $\mu$

Hypothesis Test

• **Example:** Hypothesis testing about the population mean $\mu$, at $\alpha\%$ level of significance

• $H_0$: $\mu = \mu_0$

• $H_1$: $\mu \neq \mu_0 \implies \mu = \mu_1 \neq \mu_0$

• CLT $\to \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \implies Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$

• If $H_0$ is true: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0, 1)$
  
  – $Z_0$ close to 0 $\implies$ $H_0$ probably true

• $Z_0$ “much” different from 0 $\implies$ $H_0$ probably NOT true
Statistical Inference for population average $\mu$

Hypothesis Test

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$ level of significance

- $H_0$: $\mu = \mu_0$

- $H_1$: $\mu \neq \mu_0 \implies \mu = \mu_1 \neq \mu_0$

- CLT $\rightarrow \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$ $\implies Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

- If $H_0$ is true: $Z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$
  - $Z_0$ close to 0 $\rightarrow H_0$ probably true
  - $Z_0$ "much" different from 0 $\rightarrow H_0$ probably NOT true

How "much"?
Statistical Inference for population average $\mu$

Hypothesis Test

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$ level of significance
- $H_0$: $\mu = \mu_0$
- $H_1$: $\mu \neq \mu_0 \implies \mu = \mu_1 \neq \mu_0$
- CLT $\rightarrow \bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}}) \implies Z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$

- If $H_0$ is true: $Z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$
  - $Z_0$ close to 0 $\implies H_0$ probably true
  - $Z_0$ “much” different from 0 $\implies H_0$ probably NOT true
Hypothesis Test

- Example: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  \[ H_0: \mu = \mu_0 \]
  \[ H_1: \mu \neq \mu_0 \]

  Test statistic:
  \[ Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \]
Statistical Inference for population average $\mu$

**Hypothesis Test**

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$
  
  $H_0$: $\mu = \mu_0$
  
  $H_1$: $\mu \neq \mu_0$

Test statistic: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

**Critical Points**
Statistical Inference for population average $\mu$

**Hypothesis Test**

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$
  
  $H_0: \mu = \mu_0$
  
  $H_1: \mu \neq \mu_0$
  
  **Test statistic**: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

![Standard Normal Distribution N(0,1)](image)

Rejection Region
Statistical Inference for population average $\mu$

Hypothesis Test

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  $H_0$: $\mu = \mu_0$
  $H_1$: $\mu \neq \mu_0$

  Test statistic: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

**Decision Rule:**

Reject $H_0$ if

$Z_0 < Z_{\alpha/2}$ or $Z_0 > Z_{1 - \alpha/2}$
Statistical Inference for population average $\mu$  

**Hypothesis Test**

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  - $H_0: \mu = \mu_0$
  - $H_1: \mu < \mu_0$

  **Test statistic**: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

Critical Point
Statistical Inference for population average $\mu$

**Hypothesis Test**

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  \[
  H_0 : \mu = \mu_0 \\
  H_1 : \mu < \mu_0
  \]

  **Test statistic:**

  \[
  Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}
  \]

  \[
  \text{Standard Normal Distribution } N(0,1)
  \]

  **Rejection Region**
Statistical Inference for population average $\mu$

Hypothesis Test

• Example: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

$H_0: \mu = \mu_0$
$H_1: \mu < \mu_0$

Test statistic:

$$Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

Decision Rule:

Reject $H_0$ if $Z_0 < Z_\alpha$
Statistical Inference for population average $\mu$

Hypothesis Test

- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  $H_0$: $\mu = \mu_0$
  
  $H_1$: $\mu > \mu_0$

  **Test statistic**: $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$

![Standard Normal Distribution $N(0,1)$]

**Critical Point**
Statistical Inference for population average $\mu$

**Hypothesis Test**

- Example: Hypothesis testing about the population mean $\mu$, at $\alpha\%$

  $H_0: \mu = \mu_0$
  $H_1: \mu > \mu_0$

  **Test statistic:**
  \[
  Z_0 = \frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}
  \]

  **Rejection Region**
Statistical Inference for population average $\mu$  
**Hypothesis Test**  
- **Example**: Hypothesis testing about the population mean $\mu$, at $\alpha\%$  
  
  $H_0$: $\mu=\mu_0$  
  $H_1$: $\mu > \mu_0$  

  **Test statistic**:  
  
  $$Z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$  

  **Decision Rule**:  
  Reject $H_0$ if  
  $$Z_0 > Z_{1-\alpha}$$
Statistical Inference for population average $\mu$

Key Concepts in Hypothesis Testing

- **Decision Rule:**
  - What values of the test statistic would indicate the $H_0$ is probably not supported by the observed data, hence it should be rejected.

- **P-value:**
  - The exact level of significance, i.e., the probability of observing a value as extreme or more extreme than the calculated test statistic under the null hypothesis $H_0$, e.g.,

\[
p\text{-value} = P(Z > Z_0)
\]
Statistical Inference for population average $\mu$

Steps in Hypothesis Testing

1. Set the null hypothesis $H_0$ and alternative hypothesis $H_1$
2. Set a level of significance $\alpha\%$.
3. Calculate a test statistic
4. decision rule or
5. P-value of the test statistic (preferred)
6. conclusion
We will cover examples for three cases

1) Single population: one sample t-test
   • Interested in the population mean

2) Two independent population: two sample t-test
   • Interested in comparing two population means

3) Two dependent population: Paired t test
   • Interested in comparing mean changes within subjects (before vs. after)
Statistical Inference for population average $\mu$

Case 1: One-Sample: two-sided hypothesis Test

- **Example**: We want to test the following hypothesis about the population mean $\mu$ of the systolic blood pressure of the Framingham Heart Study population, at $\alpha=5\%$ level of significance:

  $H_0: \mu = 130 \quad vs \quad H_1: \mu \neq 130$

- **Test statistic**: 
  
  $Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{136.32 - 130}{22.8 / \sqrt{11627}} = 29.91$

- **Conclusion**: $Z_0 = 29.91 \Rightarrow$ reject $H_0$ if $|Z_0| > 1.96$

- **p-value** = $P(Z>|Z_0|) = P(Z<-Z_0) + P(Z > Z_0) = 2 * P(Z>29.91) < 0.0001$

```r
t.test(dat1$sysbp, mu = 130)
> t.test(dat1$sysbp, mu = 130)

One Sample t-test

data:  dat1$sysbp
t = 29.911, df = 11626, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 130
95 percent confidence interval:  
135.9097 136.7386
sample estimates: 
mean of x
136.3241
```
Statistical Inference for population average $\mu$

One-sided hypothesis Test

- **Example**: We want to test the following hypothesis about the population mean $\mu$ of the systolic blood pressure of the Framingham Heart Study population, at $\alpha=5\%$ level of significance:

  $H_0: \mu=130$ vs $H_1: \mu > 130$

- **Test statistic**: 
  \[ Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{136.32 - 130}{22.8 / \sqrt{11627}} = 29.91 \]

- **Conclusion**: 
  $Z_0 = 29.91 \Rightarrow$ reject $H_0$ if $Z_0 > 1.68$

- **p-value** = $P(Z > Z_0) = P(Z > 29.91) < 0.0001$

```r
> t.test(dat1$sysbp, mu=130, alternative="greater")
## one-sided H1: mu > 130

One Sample t-test

data:  dat1$sysbp
  t = 29.91, df = 11626, p-value < 2.2e-16
alternative hypothesis: true mean is greater than 130
95 percent confidence interval:
   135.9763  Inf
sample estimates:
mean of x
   136.3241
```
Statistical Inference for population average $\mu$

One-sided hypothesis Test

- **Example**: We want to test the following hypothesis about the population mean $\mu$ of the systolic blood pressure of the Framingham Heart Study population, at $\alpha=5\%$ level of significance:

  \[ H_0: \mu = 130 \quad \text{vs} \quad H_1: \mu < 130 \]

- **Test statistic**:
  \[ Z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{136.32 - 130}{22.8 / \sqrt{11627}} = 29.91 \]

- **Conclusion**:
  \[ Z_0 = 29.91 \implies \text{reject } H_0: \text{if } Z_0 < -1.68 \]

- **p-value** = $P(Z < Z_0) = 1$

```r
> t.test(dat1$sysbp, mu=130, alternative="less")
# one-sided H1: mu < 130
One Sample t-test

data:  dat1$sysbp
   t = 29.911, df = 11626, p-value = 1
alternative hypothesis: true mean is less than 130
95 percent confidence interval:
   Inf 136.6719
sample estimates:
mean of x
136.3241
```
Two Independent Samples

• **Case 2**: two-independent populations (two-samples)
  - $X_1$ ‘sysbp’ of people *without previous CHD*, with $\mu_1$ and unknown $\sigma_1$
  - $X_2$ ‘sysbp’ of people *with previous CHD*, with $\mu_2$ and unknown $\sigma_2$

**Hypothesis Testing for $\mu_1 - \mu_2$**

- Null hypothesis ($H_0$): $\mu_1 - \mu_2 = 0 \implies \mu_1 = \mu_2$
- Alternative hypothesis ($H_1$):
  - $\mu_1 - \mu_2 \neq 0 \implies \mu_1 \neq \mu_2$ (two-sided test), or
  - $\mu_1 - \mu_2 < 0 \implies \mu_1 < \mu_2$ (one-sided test), or
  - $\mu_1 - \mu_2 > 0 \implies \mu_1 > \mu_2$ (one-sided test)
Two Independent Samples

• **Case 2:** two-independent populations (two-samples)
  
  – **Case 2.A: Known variances**
  
  • $X_1$ ‘sysbp’ of people **without previous CHD**, with $\mu_1$ and known $\sigma_1$
  
  • $X_2$ ‘sysbp’ of people **with previous CHD**, with $\mu_2$ and known $\sigma_2$

**Hypothesis Testing for $\mu_1-\mu_2$**

• Test statistic:
  
  $$Z_0 = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim N(0, 1)$$

• Decision Rules by $H_1$:
  
  Testing $H_0: \mu_1-\mu_2=0$ vs:

<table>
<thead>
<tr>
<th>$H_1$</th>
<th>Reject $H_0$ if:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1-\mu_2 \neq 0$</td>
<td>$Z_0 &lt; Z_{\alpha/2}$ or $Z_0 &gt; Z_{1-\alpha/2}$</td>
</tr>
<tr>
<td>$\mu_1-\mu_2 &lt; 0$</td>
<td>$Z_0 &lt; Z_\alpha$</td>
</tr>
<tr>
<td>$\mu_1-\mu_2 &gt; 0$</td>
<td>$Z_0 &gt; Z_{1-\alpha}$</td>
</tr>
</tbody>
</table>
Two Independent Samples

• **Case 2:** two-independent populations (two-samples)
  • $X_1$ ‘sysbp’ of people **without prevchd**, with $\mu_1$ and unknown $\sigma_1$
  • $X_2$ ‘sysbp’ of people **with prevchd**, with $\mu_2$ and unknown $\sigma_2$

Hypothesis Testing for $\mu_1 - \mu_2$

$H_0: \mu_1 = \mu_2$ vs. $H_1: \mu_1 \neq \mu_2$

```r
t.test(sysbp ~ prevchd, data=dat1) ## var.equal = FALSE

> t.test(sysbp ~ prevchd, data=dat1) ## var.equal = FALSE

welch Two Sample t-test
data: sysbp by prevchd
t = -13.036, df = 945.08, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-13.54697 -10.00183
sample estimates:
mean in group 0 mean in group 1
135.4714    147.2458
```
Two Dependent Samples

- **Case 3:** two-dependent populations (two-samples)
  - $X_1$ ‘sysbp’ of people at baseline, with $\mu_1$ and unknown $\sigma_1$
  - $X_2$ ‘sysbp’ of people 6yrs after baseline, with $\mu_2$ and unknown $\sigma_2$
  - Suppose variable: $d=x_1-x_2$ from population with $\mu_d$ and $\sigma_d$

  **Hypothesis Testing for $\mu_d$**

  - Null hypothesis ($H_0$): $\mu_d=0$
  - Alternative hypothesis ($H_1$):
    - $\mu_d \neq 0$ (two-sided test), or
    - $\mu_d < 0$ (one-sided test), or
    - $\mu_d > 0$ (one-sided test)

  Looks familiar? This is then same as one-sample t-test!
Power = 1 - P(Type II error) = 1 - P(do not reject $H_0$ | $H_1$ is true) = 1 - $\beta$

= P(reject $H_0$ | $H_1$ is true)

• E.g., the hypothesis: $H_0$: $\mu = \mu_0$ vs $H_1$: $\mu = \mu_1 > \mu_0$

• The power of this test is:

Power = P(reject $H_0$ | $H_1$ is true) = P( $Z_0 > Z_{1-\alpha}$ | $\mu = \mu_1 > \mu_0$ )
Power and Sample Size Determination

Power is a function of 1) standard deviation ($\sigma$),
2) sample size ($n$),
2) mean difference (or effect size),
3) type I error ($\alpha$).

The power of the test is:

\[
\text{Power} = P(\text{reject } H_0 \mid H_1 \text{ is true}) = P(Z_1 > Z_{1-\alpha} - \frac{\mu_1 - \mu_0}{\frac{\sigma}{\sqrt{n}}})
\]  

The power of the test depends on:

- \(n\) (standard deviation)
  \[\sigma \uparrow \Rightarrow \text{Power} \downarrow\]

- \(n\) (sample size)
  \[n \uparrow \Rightarrow \text{Power} \uparrow\]

- \(\alpha\) (significance level)
  \[\alpha \downarrow \Rightarrow \text{Power} \downarrow\]

- \(\mu_1 - \mu_0\) (Effect Size)
  \[\text{ES} \uparrow \Rightarrow \text{Power} \uparrow\]
Sample Size Determination

**Case 1:** Single population (one-sample):

\[ H_0: \mu = 100 \ vs \ H_1: \mu \neq 100 \]

- at \( \alpha = 5\% \) level of significance.
- We want a powerful test with power 80% power.
- The test will reject the null hypothesis if the true mean is 5 units different from 100 (either smaller or larger – two-sided test). Namely, \( |\mu - \mu_0| = 5 \).
- Suppose we know that standard deviation of the outcome variable \( \sigma = 9.5 \)
- What is the required sample size?
Sample Size Determination

**Case 1: single population (one-sample)**

\[ H_0: \mu = 100 \quad \text{vs} \quad H_1: \mu \neq 100 \]  
(two-sided test)

```r
library(pwr)
pwr.t.test(d = 5/9.5, sig.level=0.05, power = 0.8, type="one.sample")
```

```r
> pwr.t.test(d = 5/9.5, sig.level=0.05, power = 0.8, type="one.sample")
```

```
one-sample t test power calculation

n = 30.3112
d = 0.5263158
sig.level = 0.05
power = 0.8
alternative = two.sided
```

The total \( N = 31 \)
Sample Size Determination

Case 2: two dependent populations (two-samples) with unknown variance of the differences

Example: Suppose \( s_d = 7 \). We want to test the hypothesis:

\[
H_0: \mu_1 = \mu_2 = 100 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2
\]

• at \( \alpha = 5\% \) level of significance.
• We want to detect \( |\mu_1 - \mu_2| = 5 \).
• With power = 80%

What is the required sample size?
Sample Size Determination

Case 2: two dependent populations (two-samples)

$H_0: \mu_1 = \mu_2 = 100 \quad \text{vs} \quad H_1: \mu_1 \neq \mu_2$ (two-sided test)

Assume:
→ unknown variance of the differences,

i.e., $s_d = 7$

```
pwr.t.test(d = 5/7, sig.level=0.05, power = 0.8, type="two.sample")
```

```
> pwr.t.test(d = 5/7, sig.level=0.05, power = 0.8, type="two.sample")

Two-sample t test power calculation

 n = 31.75708
 d = 0.7142857
 sig.level = 0.05
 power = 0.8
 alternative = two.sided

NOTE: n is number in *each* group
```

N = 32 per group. The total N = 64.