A3. Statistical Inference

Hypothesis Testing for One Mean $\mu$

**POPULATION**

Assume random variable $X \sim N(\mu, \sigma)$.  

Testing $H_0: \mu = \mu_0$ vs. $H_A: \mu \neq \mu_0$

**ONE SAMPLE**

Test Statistic (with $s$ replacing $\sigma$ in standard error $\sigma/\sqrt{n}$):

\[
\frac{\bar{X} - \mu_0}{s/\sqrt{n}} \sim \begin{cases} 
Z, & \text{if } n \geq 30 \\
 t_{n-1}, & \text{if } n < 30
\end{cases}
\]

1a Normality can be verified empirically by checking quantiles (such as 68%, 95%, 99.7%), stemplot, normal scores plot, and/or “Lilliefors Test.” If the data turn out not to be normally distributed, things might still be OK due to the Central Limit Theorem, provided $n \geq 30$. Otherwise, a transformation of the data can sometimes restore a normal distribution.

1b When $X_1$ and $X_2$ are not close to being normally distributed (or more to the point, when their difference $X_1 - X_2$ is not), or not known, a common alternative approach in hypothesis testing is to use a “nonparametric” test, such as a Wilcoxon Test. There are two types: the “Rank Sum Test” (or “Mann-Whitney Test”) for independent samples, and the “Signed Rank Test” for paired sample data. Both use test statistics based on an ordered ranking of the data, and are free of distributional assumptions on the random variables.

2 If the sample sizes are large, the test statistic follows a standard normal $Z$-distribution (via the Central Limit Theorem), with standard error $\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. If the sample sizes are small, the test statistic does not follow an exact $t$-distribution, as in the single sample case, unless the two population variances $\sigma_1^2$ and $\sigma_2^2$ are equal. (Formally, this requires a separate test of how significantly the sample statistic $s_1^2 / s_2^2$, which follows an $F$-distribution, differs from 1. An informal rule of thumb is to accept equivariance if this ratio is between 0.25 and 4. Other, formal tests, such as “Levene’s Test”, can also be used.) In this case, the two samples can be pooled together to increase the power of the $t$-test, and the common value of their equal variances estimated. However, if the two variances cannot be assumed to be equal, then approximate $t$-tests – such as Satterwaite’s Test – should be used. Alternatively, a Wilcoxon Test is frequently used instead; see footnote 1b above.
Hypothesis Testing for Two Means $\mu_1$ vs. $\mu_2$

**POPULATION**

Random Variable $X$ defined on two groups ("arms"):

Assume $X_1 \sim N(\mu_1, \sigma_1)$, $X_2 \sim N(\mu_2, \sigma_2)$. \(^{1a,1b}\)

Testing $H_0$: $\mu_1 - \mu_2 = \mu_0$

Note: $= 0$, frequently

mean $\mu$ of $X_1 - X_2$

**TWO SAMPLES**

<table>
<thead>
<tr>
<th>Independent</th>
<th>Paired</th>
</tr>
</thead>
<tbody>
<tr>
<td>$n_1 \geq 30$, $n_2 \geq 30$</td>
<td>$n_1 \geq 30$, $n_2 &lt; 30$</td>
</tr>
<tr>
<td>Test Statistic ($\sigma_1^2, \sigma_2^2$ replaced by $s_1^2, s_2^2$ in standard error):</td>
<td>Test Statistic ($\sigma_1^2, \sigma_2^2$ replaced by $s_{\text{pooled}}^2$ in standard error):</td>
</tr>
<tr>
<td>$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \mu_0}{\sqrt{s_1^2 + s_2^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0,1)$</td>
<td>$T = \frac{(\bar{X}<em>1 - \bar{X}<em>2) - \mu_0}{\sqrt{s</em>{\text{pooled}}^2} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t</em>{df}$, $df = n_1 + n_2 - 2$</td>
</tr>
<tr>
<td>where $s_{\text{pooled}}^2 = \frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{df}$</td>
<td>where $s_{\text{pooled}}^2 = \frac{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]}{df}$</td>
</tr>
</tbody>
</table>

Since the data are naturally "matched" by design, the pairwise differences constitute a single collapsed sample. Therefore, apply the appropriate one-sample test to the random variable

$D = X_1 - X_2$

(hence $D = \bar{X}_1 - \bar{X}_2$), having mean $\mu = \mu_1 - \mu_2$; $s$ = sample standard deviation of the $D$-values.

Note that the Wilcoxon Signed Rank Test may be used as an alternative.
Hypothesis Testing for One Proportion $\pi$

**POPULATION**

Binary random variable $Y$, with $P(Y = 1) = \pi$

Testing $H_0$: $\pi = \pi_0$ vs. $H_A$: $\pi \neq \pi_0$

**ONE SAMPLE**

If $n$ is large$^3$, then standard error $\approx \sqrt{\pi(1 - \pi)/n}$ with $N(0, 1)$ distribution.

- For confidence intervals, replace $\pi$ by its point estimate $\hat{\pi} = X/n$, where $X = \sum (Y = 1) =$ # “successes” in sample.

- For acceptance regions and $p$-values, replace $\pi$ by $\pi_0$, i.e.,

$$
\text{Test Statistic: } \quad Z = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \sim N(0, 1)
$$

If $n$ is small, then the above approximation does not apply, and computations are performed directly on $X$, using the fact that it is binomially distributed. That is, $X \sim \text{Bin}(n; \pi)$. Messy by hand...

---

$^3$ In this context, “large” is somewhat subjective and open to interpretation. A typical criterion is to require that the mean number of “successes” $n\pi$, and the mean number of “failures” $n(1-\pi)$, in the sample(s) be sufficiently large, say greater than or equal to 10 or 15. (Other, less common, criteria are also used.)
Hypothesis Testing for Two Proportions $\pi_1$ vs. $\pi_2$

**POPULATION**

Binary random variable $Y$ defined on two groups ("arms"),

$P(Y_1 = 1) = \pi_1$, $P(Y_2 = 1) = \pi_2$

Testing $H_0$: $\pi_1 - \pi_2 = 0$ vs. $H_A$: $\pi_1 - \pi_2 \neq 0$

**TWO SAMPLES**

<table>
<thead>
<tr>
<th></th>
<th>Independent</th>
<th>Paired</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Large</strong></td>
<td>Standard error = $\sqrt{\pi_1 (1 - \pi_1) / n_1 + \pi_2 (1 - \pi_2) / n_2}$, with $N(0, 1)$ distribution</td>
<td>McNemar’s Test (A “matched” form of the $\chi^2$ Test.)</td>
</tr>
<tr>
<td></td>
<td>• For confidence intervals, replace $\pi_1$, $\pi_2$ by point estimates $\hat{\pi}_1$, $\hat{\pi}_2$.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>• For acceptance regions and $p$-values, replace $\pi_1$, $\pi_2$ by the pooled estimate of their common value under the null hypothesis, $\hat{\pi}_{pooled} = (X_1 + X_2) / (n_1 + n_2)$, i.e.,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test Statistic: $Z = \frac{(\hat{\pi}<em>1 - \hat{\pi}<em>2) - 0}{\sqrt{\hat{\pi}</em>{pooled} (1 - \hat{\pi}</em>{pooled}) \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Alternatively, can use a Chi-squared ($\chi^2$) Test.</td>
<td></td>
</tr>
<tr>
<td><strong>Small</strong></td>
<td>Fisher’s Exact Test (Messy; based on the “hypergeometric distribution” of $X$.)</td>
<td>Ad hoc techniques; not covered.</td>
</tr>
</tbody>
</table>