6 Hypothesis Testing Definitions and a First Test

A CLT review problem as a preview of hypothesis testing
e.g. A battery maker claims that a battery lifetime has \( \mu = 40 \) hours and \( \sigma = 5 \) hours. Suppose a random sample of 100 batteries is selected.

(a) If the claim is true, what is \( P(\bar{X} \leq 36.7) \)? (b) Based on (a), if the claim is true, is \( \bar{X} = 36.7 \) unusually short? (c) If \( \bar{X} = 36.7 \), is the claim plausible?

(a) 

(b) 

(c) 

(d) If the claim is true, what is \( P(\bar{X} \leq 39.8) \)? (e) Based on (d), if the claim is true, is \( \bar{X} = 39.8 \) unusually short? (f) If \( \bar{X} = 39.8 \), is the claim plausible?

(d) \( P(\bar{X} \leq 39.8) = P(Z < \frac{39.8 - 40}{5}) = P(Z < -0.4) = .3446 \)

(e) 

(f) 

Hypothesis testing

A hypothesis test checks whether a parameter value with a sample. It considers

- \( H_0 \), the null hypothesis, asserts, “any effect indicated by the sample is merely due to , and is an effect in the population”; and
- \( H_A \), the alternative hypothesis, which \( H_0 \), saying “the effect in the sample is in the population”

\( H_0 \) is presumed until evidence makes it unreasonable.

- Data are gathered, and a is computed from the data. The test statistic is a random variable. Its realization is evidence for deciding between \( H_0 \) and \( H_A \).
- If the test statistic is unlikely in light of \( H_0 \), we say it falls in the and we the null. Otherwise we the null.

e.g. Consider a fire alarm. The natural choices for \( H_0 \) and \( H_A \) are:

\( H_0 \): There is no fire.
\( H_A \): There is a fire.
One test statistic might be the temperature in the room: __________ is more evidence against $H_0$. Another could be the concentration of smoke particles: __________ is more evidence against the null. Let’s use temperature $T$ as our statistic.

Suppose research indicates that a temperature over 110°F indicates a fire. Then our rejection region is $T > 110$. If we measured the temperature and got $t = 70$, we would _____________. If we got $t = 200$, we would ____________ and set off the alarm.

There are four possible outcomes in a test:

<table>
<thead>
<tr>
<th>$H_0$ True</th>
<th>Reject $H_0$</th>
<th>Not Reject $H_0$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$H_0$ True</td>
<td>___________</td>
<td>type I Error / $\alpha$</td>
</tr>
<tr>
<td>$H_0$ False</td>
<td>___________</td>
<td>type II Error / $\beta$</td>
</tr>
</tbody>
</table>

The probability of a type I error is $\alpha = P(\text{Reject } H_0| H_0 \text{ is true})$. (Read “|” as __________.)

To compute $\alpha$, we need the distribution of the test statistic under $H_0$. A __________ $\alpha$ indicates a better test. In the fire alarm example, $\alpha$ is the probability of a ___________.

The probability of a type II error is $\beta = P(\text{do not reject } H_0| H_0 \text{ is false})$.

A __________ $\beta$ indicates a better test. In the fire alarm example, $\beta$ is the probability that the alarm __________ sound during a fire.

The power of a test is $= P(\text{reject } H_0| H_0 \text{ is false}) = __________$

A __________ power indicates a better test. In the fire alarm example, power is the probability that the alarm goes off __________.

We want small $\alpha$ and small $\beta$ (or equivalently, large power, $1 - \beta$). Unfortunately, for a given fixed sample size, if we adjust our rejection region to decrease $\alpha$, $\beta$ will go up, and vice versa. Thus, when choosing a rejection region, we consider the relative importance of the two types of error.

e.g. Regarding a fire alarm, which error is worse?

One way to decrease both $\alpha$ and $\beta$ is to __________.

The __________ of a test is the probability of a test statistic realizing to a value as extreme or more extreme than the one observed, when $H_0$ is true. A __________ p-value indicates more evidence against $H_0$.

The p-value required to cause a rejection of $H_0$ is called the __________ of the test. A typical level is $\alpha = ______$.

Reporting __________ is better than stating a reject or not-reject decision.
Example

Reconsider the introductory example, above:

A battery manufacturer claims that the lifetime of a type of battery has $\mu = 40$ hours and $\sigma = 5$ hours. A SRS of 100 batteries is selected.

(a) If the claim is true, $P(\bar{X} \leq 36.7) = P(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{36.7 - 40}{5/\sqrt{100}}) = P(Z < -6.6) \approx 0$

(b) Based on (a), if the claim is true, is $\bar{X} = 36.7$ unusually short? Yes.
(c) If $\bar{X} = 36.7$, is claim plausible? No.

Here it is again, using our new terminology:

- We test $H_0$: __________ vs. $H_A$: __________
- Our test statistic is $Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$
- Let’s use significance level $\alpha = .05$, so our rejection region is ________________.
- Our p-value is ________________.
- We ________________ $H_0$.
- Our conclusion is ________________.

To summarize our test,

Suppose $X_1, \ldots, X_n$ is a simple random from $N(\mu, \sigma^2)$ or $n$ is large (say $n > 30$), and $\sigma$ is known. To test that $\mu$ has a specified value, $\mu_0$,

1. State null and alternative hypotheses, $H_0 : \mu = \mu_0$ and $H_A$ (below)
2. Check assumptions
3. Find the $z$-score, $z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$
4. Find the $P$-value, which depends on $H_A$:
   - $H_A : \mu > \mu_0 \implies$ $P$-value = $P(Z > z)$, the area right of $z$
   - $H_A : \mu < \mu_0 \implies$ $P$-value = $P(Z < z)$, the area left of $z$
   - $H_A : \mu \neq \mu_0 \implies$ $P$-value = $P(|Z| > |z|)$, the sum of areas left of $-|z|$ and right of $|z|$
5. Draw a conclusion: \[\begin{cases} 
P\text{-value} \leq .05 \implies \text{reject } H_0 \\
P\text{-value} > .05 \implies \text{do not reject } H_0 \end{cases}\]
Extra example

A powdered medicine is supposed to have a mean particle diameter of $\mu = 15 \, \mu m$. Its manufacturing process is known to produce a mean particle diameter that occasionally drifts, while the standard deviation of diameters stays reasonably steady around 1.8 $\mu m$. A simple random sample of 87 particles had a mean diameter of 15.2 $\mu m$. Is this strong evidence that the powder does not meet its specification? (If yes, the manufacturing process needs to be recalibrated.)

- **Hypotheses:**

- **Assumptions:**

- **Test statistic:**

- **P-value:**

- **Conclusion:**

In the next section, we consider other tests, each based on a sample from ______ population.