Statistical Hypothesis Testing
Statistical Hypothesis Testing

- definitions
  - what is it?
- theories
  - Why are we using it?
- Examples
  - How are we using it?
- Applications
  - How is it related to us?
What is it?

A statistical test provides a mechanism for making quantitative decisions about a process or processes. The intent is to determine whether there is enough evidence to "reject" a conjecture or hypothesis about the process.

Statistical Hypothesis Testing

- Statistical
- Hypothesis
- Testing
Statistical Hypothesis Testing

- statistical model & statistical inference
  - random variables $X$
  - distribution $P_{\Theta}$ (at least partly unknown)
  - a set of observations/measurement
Statistical Hypothesis Testing

- statistical model & statistical inference
  - “Statistical inference is concerned with methods of using this observational material to obtain information concerning the distribution of $X$ or the parameter $\Theta$ with which it is labeled”
  - “A statistical inference is a procedure that produces a probabilistic statement about … a statistic model”

Lehmann, E.L. Testing Statistical Hypotheses
DeGroot M.H., Schervish M.J. Probability and Statistics
Statistical Hypothesis Testing

- Sample vs. Population
- Example ("Lake Wobegon")
  - Someone claims kids at Lake Wobegon have above average intelligence
  - random samples of 9 kids there with test result of {116, 128, 125, 119, 89, 99, 105, 116, 118}
  - Wechsler scores (the test they take) are scaled to be normally distributed with a mean of 100 and standard deviation of 15.
    - mean of sample: 112.8

http://www.sjc-su.edu/faculty/gerstman/StaPrimer/hyp-test.pdf
Statistical Hypothesis Testing

- Null Hypothesis vs. Alternative Hypothesis
- Null hypothesis being “attacked”
- Usually more emphasis on Alternative Hypothesis
  - “Formulate the null hypothesis $H_0$ (commonly, that the observations are the result of pure chance) and the alternative hypothesis $H_\alpha$ (commonly, that the observations show a real effect combined with a component of chance variation).”

Statistical Hypothesis Testing

- The idea is to refute $H_0$ if the sample is statistically far from the population/distribution, which is modeled only under the assumption that $H_0$ is true.
- We want to know the confidence level of refuting $H_0$.
- Failed to reject the null hypothesis does not mean the null hypothesis is true.
Statistical Hypothesis Testing

- one-sided vs. two-sided alternatives
- one-sided:
  - $H_0$: $u \leq 100$
  - $H_a$: $u > 100$
- two-sided:
  - $H_0$: $u = 100$
  - $H_a$: $u \neq 100$

Statistical Hypothesis Testing

Statistical Hypothesis Testing

Statistical Hypothesis Testing

- **Test Statistic:**
  - The test statistic is a statistical method based on the specific hypothesis test.
  - $T = r(X)$ where $X$ is the random sample from the distribution.
  - $H_0$ will be rejected if $T \in \mathbb{R}$

- **Critical Region:**
  - The set $S_1 = \{ x : r(x) \in \mathbb{R} \}$
Statistical Hypothesis Testing

- **Significance Level:** $\alpha$
  - “A value of $\alpha = 0.05$ means that we inadvertently reject the null hypothesis 5% of the time when it is in fact true.”

- **Power:** $1 - \beta$
  - “the probability of accepting the null hypothesis when the alternative hypothesis is, in fact, true, is called $\beta$.”

- often referred as Type 1 and Type 2 errors, respectively

NIST/SEMATECH e-Handbook of Statistical Methods
Statistical Hypothesis Testing

- Critical Region: encompasses those values of the test statistic that lead to a rejection of the null hypothesis.
- P-value: the probability that a test statistic at least as significant as the one observed
Statistical Hypothesis Testing

Two-Tailed Test Critical Value = ± 1.9723

Upper-Tailed Test Critical Value = 1.6527

Lower-Tailed Test Critical Value = -1.6527
Why are we using it? What’s an alternative?

- Quantitative understanding of data
- Exploratory Data Analysis (EDA)
  - The primary goal of EDA is to maximize the analyst's insight into a data set and into the underlying structure of a data set
- Quantitative (Classic) Techniques
  - Hypothesis tests
  - Interval estimation
    - An interval estimate quantifies … uncertainty in the sample estimate by computing lower and upper values of an interval

NIST/SEMATECH e-Handbook of Statistical Methods
Z-statistics vs. T-statistics

Z-statistic:
\[ Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \]
- Normally distributed
- Ok if \( n > 30 \)

T-statistic:
\[ t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \]
- T-distributed
- Ok if small
Example

- Someone claims kids at Lake Wobegon have above average intelligence
- random samples of 9 kids there with test result of \{116, 128, 125, 119, 89, 99, 105, 116, 118\}
- Wechsler scores (the test they take) are scaled to be normally distributed with a mean of 100 and standard deviation of 15.
- mean of sample: 112.8

Example

- mean(X) = 112.8
- u = 100
- test statistic: z-statistic
- H₀: u = 100
- Hₐ: u > 100

\[ z_{stat} = \frac{\bar{x} - \mu_0}{SEM} \]

- SEM = σ/sqrt(N) = 15/sqrt(9) = 5
- \[ Z_{stat} = \frac{(112.8 - 100)}{5} = 2.56 \]
- look up Z table → p = 0.0052 = 0.52%

Example

Standard Normal Probabilities

Table entry for \( z \) is the area under the standard normal curve to the left of \( z \).

<table>
<thead>
<tr>
<th>( z )</th>
<th>.00</th>
<th>.01</th>
<th>.02</th>
<th>.03</th>
<th>.04</th>
<th>.05</th>
<th>.06</th>
<th>.07</th>
<th>.08</th>
<th>.09</th>
</tr>
</thead>
<tbody>
<tr>
<td>-3.3</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0004</td>
<td>.0003</td>
</tr>
<tr>
<td>-3.2</td>
<td>.0007</td>
<td>.0007</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0006</td>
<td>.0005</td>
<td>.0005</td>
<td>.0005</td>
</tr>
<tr>
<td>-3.1</td>
<td>.0010</td>
<td>.0009</td>
<td>.0009</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0008</td>
<td>.0007</td>
<td>.0007</td>
</tr>
<tr>
<td>-3.0</td>
<td>.0013</td>
<td>.0013</td>
<td>.0012</td>
<td>.0012</td>
<td>.0011</td>
<td>.0011</td>
<td>.0011</td>
<td>.0010</td>
<td>.0010</td>
<td>.0010</td>
</tr>
<tr>
<td>-2.9</td>
<td>.0019</td>
<td>.0018</td>
<td>.0017</td>
<td>.0016</td>
<td>.0016</td>
<td>.0015</td>
<td>.0015</td>
<td>.0014</td>
<td>.0014</td>
<td>.0014</td>
</tr>
<tr>
<td>-2.8</td>
<td>.0026</td>
<td>.0025</td>
<td>.0024</td>
<td>.0023</td>
<td>.0023</td>
<td>.0022</td>
<td>.0021</td>
<td>.0021</td>
<td>.0020</td>
<td>.0019</td>
</tr>
<tr>
<td>-2.7</td>
<td>.0035</td>
<td>.0034</td>
<td>.0033</td>
<td>.0032</td>
<td>.0031</td>
<td>.0030</td>
<td>.0029</td>
<td>.0028</td>
<td>.0027</td>
<td>.0026</td>
</tr>
<tr>
<td>-2.6</td>
<td>.0047</td>
<td>.0045</td>
<td>.0044</td>
<td>.0043</td>
<td>.0041</td>
<td>.0040</td>
<td>.0039</td>
<td>.0038</td>
<td>.0037</td>
<td>.0036</td>
</tr>
<tr>
<td>-2.5</td>
<td>.0062</td>
<td>.0060</td>
<td>.0059</td>
<td>.0057</td>
<td>.0055</td>
<td>.0054</td>
<td>.0052</td>
<td>.0051</td>
<td>.0049</td>
<td>.0048</td>
</tr>
<tr>
<td>-2.4</td>
<td>.0082</td>
<td>.0080</td>
<td>.0078</td>
<td>.0075</td>
<td>.0073</td>
<td>.0071</td>
<td>.0069</td>
<td>.0068</td>
<td>.0066</td>
<td>.0064</td>
</tr>
<tr>
<td>-2.3</td>
<td>.0107</td>
<td>.0104</td>
<td>.0102</td>
<td>.0099</td>
<td>.0096</td>
<td>.0094</td>
<td>.0091</td>
<td>.0089</td>
<td>.0087</td>
<td>.0084</td>
</tr>
<tr>
<td>-2.2</td>
<td>.0139</td>
<td>.0136</td>
<td>.0132</td>
<td>.0129</td>
<td>.0125</td>
<td>.0122</td>
<td>.0119</td>
<td>.0116</td>
<td>.0113</td>
<td>.0110</td>
</tr>
<tr>
<td>-2.1</td>
<td>.0179</td>
<td>.0174</td>
<td>.0170</td>
<td>.0166</td>
<td>.0162</td>
<td>.0158</td>
<td>.0154</td>
<td>.0150</td>
<td>.0146</td>
<td>.0143</td>
</tr>
<tr>
<td>-2.0</td>
<td>.0228</td>
<td>.0222</td>
<td>.0217</td>
<td>.0212</td>
<td>.0207</td>
<td>.0202</td>
<td>.0197</td>
<td>.0192</td>
<td>.0188</td>
<td>.0183</td>
</tr>
<tr>
<td>-1.9</td>
<td>.0287</td>
<td>.0281</td>
<td>.0274</td>
<td>.0268</td>
<td>.0262</td>
<td>.0256</td>
<td>.0250</td>
<td>.0244</td>
<td>.0239</td>
<td>.0233</td>
</tr>
</tbody>
</table>

http://www.stat.ufl.edu/~athienit/Tables/Ztable.pdf
Example

Example

- Define conventions:
  - When $p$ value $> .10$ → the observed difference is “not significant”
  - When $p$ value $\leq .10$ → the observed difference is “marginally significant”
  - When $p$ value $\leq .05$ → the observed difference is “significant”
  - When $p$ value $\leq .01$ → the observed difference is “highly significant”

How is it related to us?

- Test Calibration
- Test Constraints
- QA audit
- Actually test some hypothesis…
How can we use it?

- **Data Acquisition:**
  - **Input side:**
    - A sensor input value
  - **Output side:**
    - VSCADA
    - raw sensor data is retrievable and transformable

- **Post analysis:**
  - apply suitable test statistic using favorable tools