Sharpening Our Statistical Toolkit

A Rebirth of Classical Powerful Techniques

[Image of two portraits]

Austin Menger
austin.menger@gmail.com
A Bit About Me

Current
Statistics PhD Student, University of Connecticut

Previous
New York Engineers - Lead Data Scientist
Reed Exhibitions – Database Analyst

Education
Columbia University – M.S. Applied Analytics
Georgetown University – B.S (Hons.) Mathematics

Interests
Rugby
Freestyle Skiing
## Agenda

<table>
<thead>
<tr>
<th>Sections</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Contextualizing Hypothesis Testing</strong></td>
<td></td>
</tr>
<tr>
<td>What is a Data Scientist?</td>
<td>4</td>
</tr>
<tr>
<td>Where We Are In The Process</td>
<td>5,6</td>
</tr>
<tr>
<td>Describing Data - Probabalistic Distributions</td>
<td>7-10</td>
</tr>
<tr>
<td>Hypothesis Testing Theory</td>
<td>11,12</td>
</tr>
<tr>
<td><strong>1 Sample Inference</strong></td>
<td></td>
</tr>
<tr>
<td>1-Sample Z-tests, T-tests</td>
<td>13-17</td>
</tr>
<tr>
<td>1-Sample Nonparametric Approach</td>
<td>18,19</td>
</tr>
<tr>
<td><strong>2 Sample Inference</strong></td>
<td></td>
</tr>
<tr>
<td>Welch-Satterthwaite T-test (2-Sample T-Test)</td>
<td>20</td>
</tr>
<tr>
<td>Pooled &amp; Paired T-tests</td>
<td>21-24</td>
</tr>
<tr>
<td>Business Application (Lead Scoring)</td>
<td>25-29</td>
</tr>
<tr>
<td><strong>Further Extension - ANOVA</strong></td>
<td>30</td>
</tr>
</tbody>
</table>
What is a Data Scientist?

What We Are Not

Statistics & Mathematics

Data Mining

Communication

What We Are:

Statistics & Mathematics
Where We Are In The Process

Problem Agreement

Gaps Analysis

EDA/Statistical Modeling

Change Management**
Where We Are In The Process

1. Problem Agreement
2. Gaps Analysis
3. EDA/Statistical Modeling
4. Change Management**
Describing Data – Probabilistic Distributions

General Properties of Distributions

1) \[ \sum_{i=1}^{n} P(X_i) = 1 \]

2) \[ 0 \leq P(X_i) \leq 1 \forall i, i = 1, \ldots, n \]

Example:

\[ f(x) = \frac{x}{5} \quad x = 1, 3 \]

Function (Outputs Probability)

Support
### Describing Data – Probabilistic Distributions

#### Discrete vs Continuous Distributions:

<table>
<thead>
<tr>
<th></th>
<th>Discrete</th>
<th>Continuous</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Definition</strong></td>
<td>Takes specific values</td>
<td>Takes values in an interval</td>
</tr>
<tr>
<td><strong>Support</strong></td>
<td>$x=1,3$</td>
<td>$1 \leq x \leq 3$</td>
</tr>
<tr>
<td></td>
<td>$x=0,1,2,3,...$</td>
<td>$\infty &lt; x &lt; \infty$ (x $\in$ R)</td>
</tr>
<tr>
<td><strong>Finding Probabilities</strong></td>
<td>Sum the probabilities of</td>
<td>Take the area under the curve between</td>
</tr>
<tr>
<td></td>
<td>the x values satisfying</td>
<td>the two points</td>
</tr>
<tr>
<td></td>
<td>the inequality</td>
<td></td>
</tr>
<tr>
<td><strong>Examples</strong></td>
<td>- Previous Slide</td>
<td>- Normal Dist’n</td>
</tr>
<tr>
<td></td>
<td>- Binomial Dist’n</td>
<td>- T Dist’n</td>
</tr>
<tr>
<td></td>
<td></td>
<td>- Uniform Distribution</td>
</tr>
</tbody>
</table>
Examples of Symmetric Distributions:

**Normal Distribution**

\[ X \sim N(\mu, \sigma^2) \]

\[ f(x) = \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad x \in \mathbb{R} \]

**t Distribution**

\[ X \sim t_\nu \]

\[ f(x) = \frac{\Gamma\left(\frac{\nu + 1}{2}\right)}{\sqrt{\nu \pi} \Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu + 1}{2}} \quad x \in \mathbb{R} \]
Describing Data – Probabilistic Distributions

Why Normal and T-Distributions?

Normal Dist’n
> pnorm(value, mean, st dev)
> qnorm(prob in dec form, mean, st dev)

T-Dist’n
> pt(value, degrees of freedom)
> qt(prob in dec form, d.f.)
Hypothesis Testing Theory

Conclusion: If probability “low”, yet we did observe it, then null hypothesis cannot be true

Calculation: Calculate theoretical probability of observing what you did

Observation: Observe Real Data

Base Assumption: The null hypothesis is true
Hypothesis Testing Theory - Components

1. Null and alternative hypothesis
   \(H_0\): typically the “status quo” (null)
   \(H_A\): what you’d like to test (alt)

2. Observed Test Statistic – calculated using test-specific formula (usually \(t\) (t-dist’n) or \(z\) (normal dist’n))

3. Decision Rule – Based on p-value (the probability of observing the data you did, or more extreme, given that the null hypothesis is true)
   - P-value < .05 \(\rightarrow\) Reject \(H_0\), conclude \(H_A\)
   - P-value > .05 \(\rightarrow\) Cannot reject \(H_0\), therefore cannot conclude \(H_A\)

4. Conclude in context
1 Sample Inference – 1-Sample t and z tests

### 1-Sample z-test

**Assumptions:**
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown $\mu$ (mean)
- Known $\sigma^2$

$H_0: \mu = 5 (\mu_0)$  
$H_A: \mu \neq 5$
- $\mu > 5$
- $\mu < 5$

**Test Stat:**  
$$z^* = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{\sigma} \sim N(0, 1)$$

**Rejection Rule (in R):**
- $1 - \text{pnorm}(z^*, 0, 1)$ if $\mu > \mu_0$
- $\text{pnorm}(z^*, 0, 1)$ if $\mu < \mu_0$
- $2(1 - \text{pnorm}(|z^*|, 0, 1))$ if $\mu \neq \mu_0$

### 1-Sample t-test

**Assumptions:**
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown $\mu$ (mean)
- Unknown $\sigma^2$

$H_0: \mu = 5 (\mu_0)$  
$H_A: \mu \neq 5$
- $\mu > 5$
- $\mu < 5$

**Test Stat:**  
$$t^* = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} \sim t_{n-1}$$

**Rejection Rule (in R):**
- $1 - \text{pt}(t^*, n-1)$ if $\mu > \mu_0$
- $\text{pt}(t^*, n-1)$ if $\mu < \mu_0$
- $2(1 - \text{pt}(|t^*|, 0, 1))$ if $\mu \neq \mu_0$
Application – Quality Assurance

Philips produces 65W Dimmable LED Energy Star Light Bulbs sold at Home Depot. On the Home Depot site, they advertise the “life hours” of each light bulb is 25000.

Question of Interest: Accounting for variability, is the mean lifetime of light bulbs actually 25000?
1 Sample Inference – 1-Sample t and z tests

Information Needed for the test:
- Sample of reasonable size, observing the actual lifetimes of lightbulbs in controlled environment
- Either we can use the known standard deviation over time of all light bulbs (if we have it) or just use the sample standard deviation

Assume we have a sample of n=100 light bulbs with $\bar{x} = 23024$ and sample st dev (s) = 6705

Step 1: Confirm Assumptions
- One sample compared to known value
- Testing for true unknown mean ($\mu$)
- In this case we have unknown $\sigma^2$
- Do we have approximate normality??
Confirming Normality

**R Code**

```r
qqnorm(data)
qqline(data, col="red")
```
1 Sample Inference – 1-Sample t and z tests

Running the Test

1-Sample t-test

Assumptions:
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown $\mu$ (mean)
- Unknown $\sigma^2$

$H_0: \mu = 25000 \ (\mu_0)$
$H_A: \mu < 25000$

Test Stat: $t^* = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} = \frac{\sqrt{100(23024 - 25000)}}{6705} = -2.95$

Rejection Rule (in R):

1. $1 - pt(t^*, n - 1)$ if $\mu > \mu_0$
2. $pt(t^*, n - 1)$ if $\mu < \mu_0$
3. $2(1 - pt(|t^*|, 0, 1))$ if $\mu \neq \mu_0$

Conclusion

At the alpha=.05 significance level, with a p-value=.002 < .05, we can reject the null hypothesis and conclude that the average lifetime of lightbulbs produced is shorter than the claimed 25000 hours.

$pt(-2.95, 99)=.002$
What Happens When We Don’t Have Normality?
1 Sample Inference – Nonparametric Approach

Conclusion

At the alpha=.05 significance level, with a p-value=0.1933>.05, we cannot reject the null hypothesis and therefore cannot conclude that the median lifetime of lightbulbs produced is shorter than the claimed 25000 hours.

Wilcoxon Signed Rank Test

Assumptions:
- One sample compared to known value
- Unknown $m$ (median)
- Symmetric (look at histogram)

$H_0: m = 25000 \ (m_0)$
$H_A: m \neq 25000$
- $m > 25000$
- $m < 25000$

Test Stat: Sum the positive-signed ranks \(V\)

Rejection Rule (in R):
- $1 - \text{pnorm}(z^*, 0, 1)$ if $m > m_0$
- $\text{pnorm}(z^*, 0, 1)$ if $m < m_0$
- $2(1 - \text{pnorm}(|z^*|, 0, 1))$ if $m \neq m_0$

R code:

```r
> wilcox.test(lightbulb, alternative="less", mu=25000, conf.level=.95, exact=FALSE)

Wilcoxon signed rank test with continuity correction
data: lightbulb
V = 73, p-value = 0.1933
alternative hypothesis: true location is less than 25000
```
Welch-Satterthwaite (2-Sample T-test)

**Assumptions:**
- Comparing means of two independent samples
  - Each sample approx. normal ($qqnorm$ in R)
- unknown $\sigma_1^2 \neq \sigma_2^2$

\[ H_0: \mu_1 = \mu_2 \ (\mu_1 - \mu_2 = 0) \]
\[ H_A: \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0) \]
\[ \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0) \]
\[ \mu_1 < \mu_2 \ (\mu_1 - \mu_2 < 0) \]

**Test Stat:**
\[
t^* = \frac{Y_1 - Y_2}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}}
\]

**Rejection Rule (in R):**
\[ 1 - pt(t^*, 0, 1) \quad \text{if} \ \mu_1 > \mu_2 \]
\[ pt(t^*, 0, 1) \quad \text{if} \ \mu_1 < \mu_2 \]
\[ 2(1 - pt(|t^*|, v)) \quad \text{if} \ \mu_1 \neq \mu_2 \]

Where \( v = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1-1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2-1}} = \text{degrees of freedom} \)
2 Sample Inference – Pooled & Paired T-tests

Pooled T-test

Assumptions:
- Comparing means of two independent samples
- Each sample approx normal (qqnorm in R)
- Unknown $\sigma_1^2 = \sigma_2^2$

$H_0: \mu_1 = \mu_2 \ (\mu_1 - \mu_2 = 0)$
$H_A: \mu_1 \neq \mu_2 \ (\mu_1 - \mu_2 \neq 0)$

Test Stat: $t^* = \frac{\bar{Y}_1 - \bar{Y}_2}{\sqrt{(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2})(\frac{1}{n_1} + \frac{1}{n_2})}}$

Rejection Rule (in R):
1. $1 - pt(t^*, n_1 + n_2 - 2)$ if $\mu_1 > \mu_2$
2. $pt(t^*, n_1 + n_2 - 2)$ if $\mu_1 < \mu_2$
3. $2(1 - pt(|t^*|, n_1 + n_2 - 2))$ if $\mu_1 \neq \mu_2$

Paired t-test

Assumptions:
- Comparing means of dependent samples
- Each sample approx normal (qqnorm in R)
- Unknown $\sigma^2$

$H_0: \mu_1 - \mu_2 = 0 \ (\text{interested in difference})$
$H_A: \mu_1 - \mu_2 \neq 0$

$H_0: \mu_1 - \mu_2 = 0$
$H_A: \mu_1 - \mu_2 \neq 0$

Test Stat: $t^* = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} \sim t_{n-1}$

Rejection Rule (in R):
1. $1 - pt(t^*, n - 1)$ if $\mu_1 - \mu_2 > 0$
2. $pt(t^*, n - 1)$ if $\mu_1 - \mu_2 < 0$
3. $2(1 - pt(|t^*|, 0, 1))$ if $\mu_1 - \mu_2 \neq 0$
2 Sample Inference – Pooled & Paired T-tests

R CODE

**Welch=Satterthwaite (2-Sample T-Test)**

```r
> t.test(data$response~data$group, alternative="less", paired=FALSE, var.equal=FALSE)

Welch Two Sample t-test

data:  data$response by data$group
t = -0.28737, df = 3.1286, p-value = 0.3959
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 18.81038
sample estimates:
mean in group 1 mean in group 2
 28.33333       31.00000
```

**Paired T-Test**

```r
> t.test(data$response~data$group, alternative="less", paired=TRUE, var.equal=FALSE)

Paired t-test

data:  data$response by data$group
t = -0.55074, df = 2, p-value = 0.3186
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
  -Inf 11.47175
sample estimates:
mean of the differences
  -2.666667
```
2 Sample Inference – Pooled & Paired T-tests

R CODE

Welch=Satterthwaite (2-Sample T-Test)

```R
> t.test(data$response~data$group, alternative="less", paired=FALSE, var.equal=FALSE)

Welch Two Sample t-test

data:  data$response by data$group
t = -0.28737, df = 3.1286, p-value = 0.3959
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
Inf 18.81038
sample estimates:
mean in group 1 mean in group 2
28.33333 31.00000
```

Pooled T-Test

```R
> t.test(data$response~data$group, alternative="less", paired=FALSE, var.equal=TRUE)

Two Sample t-test

data:  data$response by data$group
t = -0.28737, df = 4, p-value = 0.3941
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
Inf 17.11603
sample estimates:
mean in group 1 mean in group 2
28.33333 31.00000
```
Testing For Equal Variance

$H_0$: variances equal
$H_A$: variances not equal

```r
> var.test(data$response~data$group, alternative="two.sided")

F test to compare two variances

data:  data$response by data$group
F = 3.235, num df = 2, denom df = 2, p-value = 0.4723
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.08294802 126.16393443
sample estimates:
  ratio of variances
  3.234973
```

Conclusion
At the alpha=.05 significance level, with a p-value=.4723>.05, we conclude variances are equal.
Application – Lead Source Comparison

Projects come from various lead sources. Here, we are interested in comparing 2 lead sources that the sales team can’t agree on as the company’s “best” lead source.

Information We Have

- Project revenue for each lead noting the source each lead came from (8 sources in total)
- We consider all the history we have for each of the two lead sources we are interested in comparing as separate samples
- We gather sample statistics from each of the two lead sources

Source #1 – Think! Architecture

\[ n_1 = 293 \]
\[ \bar{x}_1 = \$76,725 \]
\[ s_1 = \$9,673 \]

Source #2 – Superstructures

\[ n_2 = 290 \]
\[ \bar{x}_2 = \$78,547 \]
\[ s_2 = \$8,431 \]
Thought Flow

1 or 2 Samples?  
Step 1

Independent or Dependent Samples?  
Step 2

Unknown $\sigma_1^2, \sigma_2^2$?  
Step 3

Equal Variance?  
Step 4

Both samples approx. Normal?  
Step 5

2 Samples (Interested in Comparing Means)  
Independent (Design Setup)  
Yes, this is typical  
Equality of Variance Test  
Normal QQ Plots
2 Sample Inference – Business Application (Lead Scoring)

```r
> var.test(data$response ~ data$group, alternative="two.sided")

F test to compare two variances

data:  data$response by data$group
F = 3.235, num df = 2, denom df = 2, p-value = 0.4723
alternative hypothesis: true ratio of variances is not equal to 1
95 percent confidence interval:
  0.08294802 126.16393443
sample estimates:
  ratio of variances
          3.234973
```

Conclusions

- Variances are equal
- Normality is satisfied
2 Sample Inference – Business Application (Lead Scoring)

**Paired t-test**

*Assumptions:*
- Compare means of DEPENDENT samples
- Each sample approx normal (qqnorm in R)
- Unknown $\sigma^2$

**1-Sample t-test**

*Assumptions:*
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown $\mu$ (mean)
- Unknown $\sigma^2$

**1-Sample z-test**

*Assumptions:*
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown $\mu$ (mean)
- Known $\sigma^2$

**Data Description**

- 2 samples comparison of means
- Independent samples
- Unknown $\sigma_1^2, \sigma_2^2$
- Equal variance
- Normality Satisfied

**Welch-Satterthwaite (2-Sample T-test)**

*Assumptions:*
- Comparing means of two independent samples
- Each sample approx. normal (qqnorm in R)
- Unknown $\sigma_1^2 \neq \sigma_2^2$

**Pooled T-test**

*Assumptions:*
- Comparing means of two indep samples
- Each sample approx normal (qqnorm in R)
- Unknown $\sigma_1^2 = \sigma_2^2$
2 Sample Inference – Business Application (Lead Scoring)

Pooled T-test

\[ H_0: \mu_1 = \mu_2 \quad (\mu_1 - \mu_2 = 0) \]
\[ H_A: \mu_1 > \mu_2 \quad (\mu_1 - \mu_2 > 0) \]

Where Group1=Think Architecture
Group 2=Superstructures

Conclusion

At the alpha=.05 significance level, with a p-value=.6589>.05, we cannot reject the null hypothesis and therefore cannot conclude that the mean revenue from Think! Architecture is greater than that from Superstructure.
Further Extension - ANOVA

What Happens When We Want to Compare 3 or more Groups?

Sample 1
\[ x_1 \]
\[ x_2 \]
\[ x_3 \]
\[ \vdots \]
\[ x_n \]

Sample 2
\[ y_1 \]
\[ y_2 \]
\[ y_3 \]
\[ \vdots \]
\[ y_n \]

Sample 3
\[ z_1 \]
\[ z_2 \]
\[ z_3 \]
\[ \vdots \]
\[ z_n \]

Analysis of Variance (ANOVA)

Resource:
https://onlinecourses.science.psu.edu/stat502/
Thank You!

Questions or Comments?