Sharpening Our Statistical Toolkit

A Rebirth of Classical Powerful Techniques

Austin Menger	austin.menger@gmail.com

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

A Bit About Me

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Current

Statistics PhD Student, University of Connecticut

Previous

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

Eduction

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

Interests

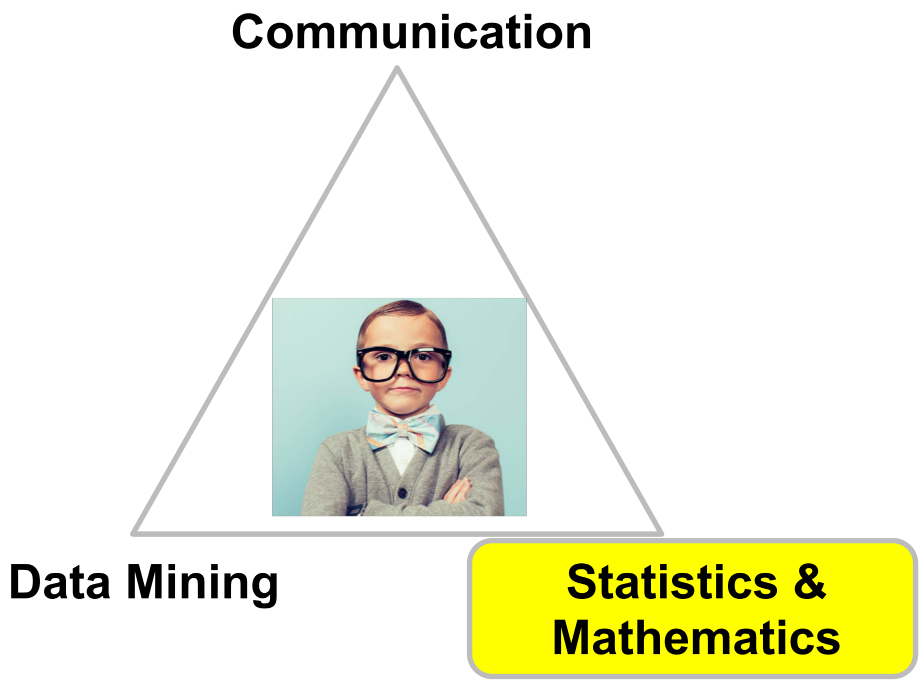
Rugby
Freestyle Skiing

Agenda

What is a Data Scientist?

What We Are Not

What We Are:



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Where We Are In The Process

Problem Agreement

↓

Gaps Analysis

↓

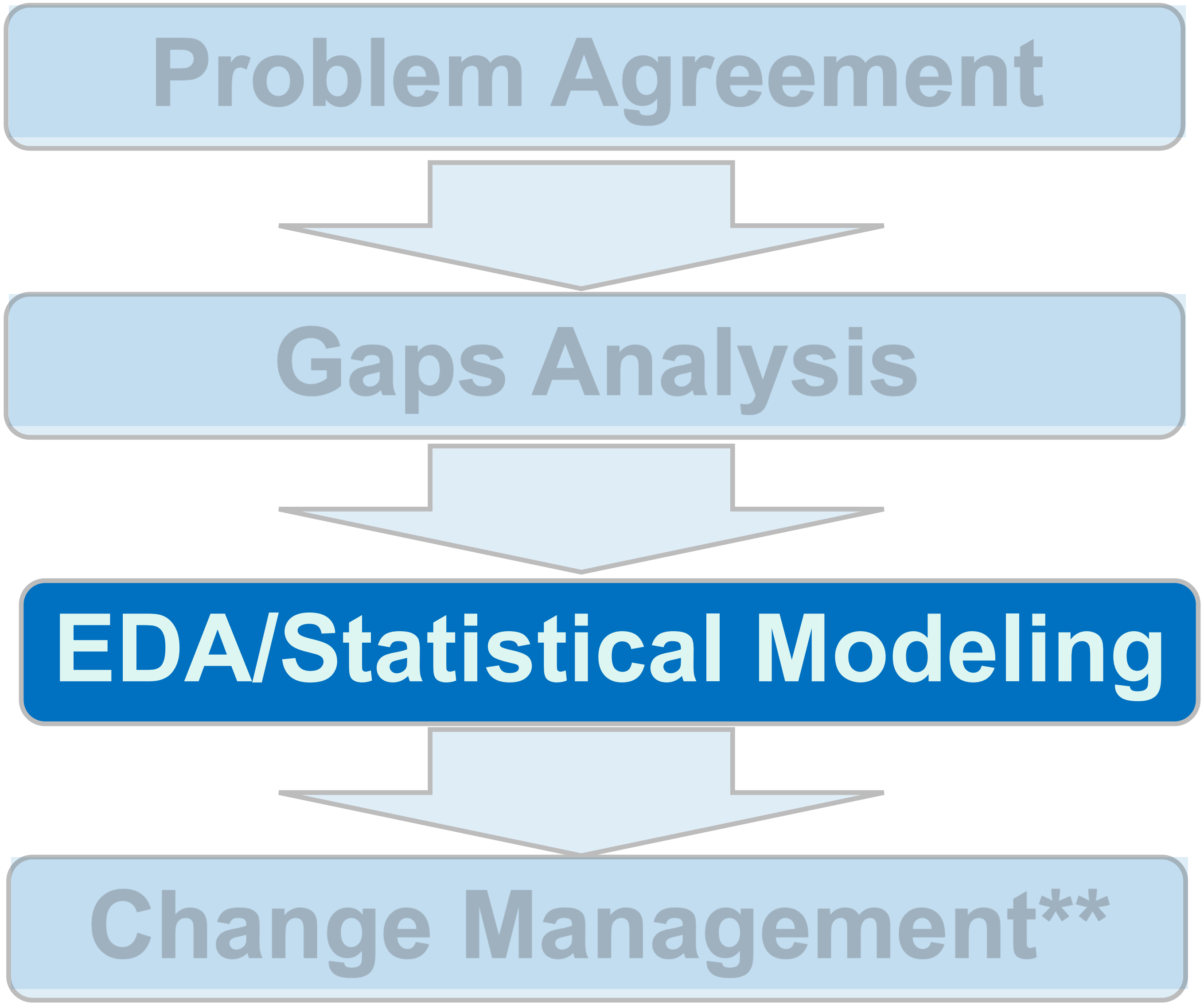
EDA/Statistical Modeling

↓

Change Management**

Columbia University
In the City of New York

Where We Are In The Process



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

Describing Data – Probabilistic Distributions

General Properties of Distributions

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

2) $$0 \leq P(X_i) \leq 1 \; \forall i, i = 1, \dots, n$$

Example:

$$f(x) = \frac{x}{5} \qquad x = 1,3$$

↑ **Function (Outputs Probability)**

↑ **Support**

Describing Data – Probabilistic Distributions

Discrete vs Continuous Distributions:

	Discrete	Continuous
Definition	Takes specific values	Takes values in an interval
Support	x=1,3 x=0,1,2,3,…	$1 \leq x \leq 3$ $\infty < x < \infty$ ($x \in R$)
Finding Probabilities	Sum the probabilities of the x values satisfying the inequality	Take the area under the curve between the two points
Examples	- Previous Slide - Binomial Dist'n	- Normal Dist'n - T Dist'n - Uniform Distribution

Hypothesis Tests Utilize Continuous Distributions

Describing Data – Probabilistic Distributions

Examples of Symmetric Distributions:

Normal Distribution	$X \sim N(\mu, \sigma^2)$	$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$	$x \epsilon \mathbb{R}$
t Distribution	$X \sim t_\nu$	$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$	$x \epsilon \mathbb{R}$

Describing Data – Probabilistic Distributions

Why Normal and T-Distributions?

Standard Normal Probabilities

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

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

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 → Reject H_0, conclude H_A
P-value >.05 → Cannot reject H_0, therefor 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 μ (mean)
- Known σ^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):

$1 - pnorm(z^*, 0, 1) \quad if \quad \mu > \mu_0$

$pnorm(z^*, 0, 1) \quad if \quad \mu < \mu_0$

$2(1 - pnorm(|z^*|, 0, 1) \quad if \quad \mu \neq \mu_0$

1-Sample t-test

Assumptions:

-One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown μ (mean)
- Unknown σ^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):

$1 - pt(t^*, n-1) \quad if \quad \mu > \mu_0$

$pt(t^*, n-1) \quad if \quad \mu < \mu_0$

$2(1 - pt(|t^*|, 0, 1) \quad if \quad \mu \neq \mu_0$

1 Sample Inference – 1-Sample t and z tests

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.

Philips Model # 465996 ★★★★★ (12)

65W Equivalent Soft White with Warm Glow BR30 Dimmable LED Ene...

$13.97

TOP

Product Overview | Specifications | Questions & Answers | Customer Reviews

Product Overview

Energy Star Certified and unlike standard LED's, these Philips bulbs offer a dimmable warm glow effect that lets you go from functional lighting, to inviting, to cozy. You can customize your room for every moment and always have the right light. Perfect for indoor track fixtures, down lights and high hats to create a lovely, warm ambiance.

- Brightness: 650-Lumens
- Estimated yearly energy cost: $1.07 (based on 3-hours/day, 11¢/kWh, cost depend on rates and use)
- Life hours: 25000
- Light appearance: soft white
- Energy used: 9-Watt
- Lumens per watt: 72.22
- Enjoy the energy-savings of LED's without sacrificing light quality with this warm glow dimmable bulb
- Ideal for indoor use in track fixtures, high hats and down lights in living rooms, bedrooms, dining and family rooms
- With a lifetime of up to 25,000-hours, you can reduce the hassle of frequently replacing your light bulbs, Philips LED bulbs enable the perfect lighting solution for 22+ years
- Just by flipping the switch, your room is at full brightness, no slow starting or waiting

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

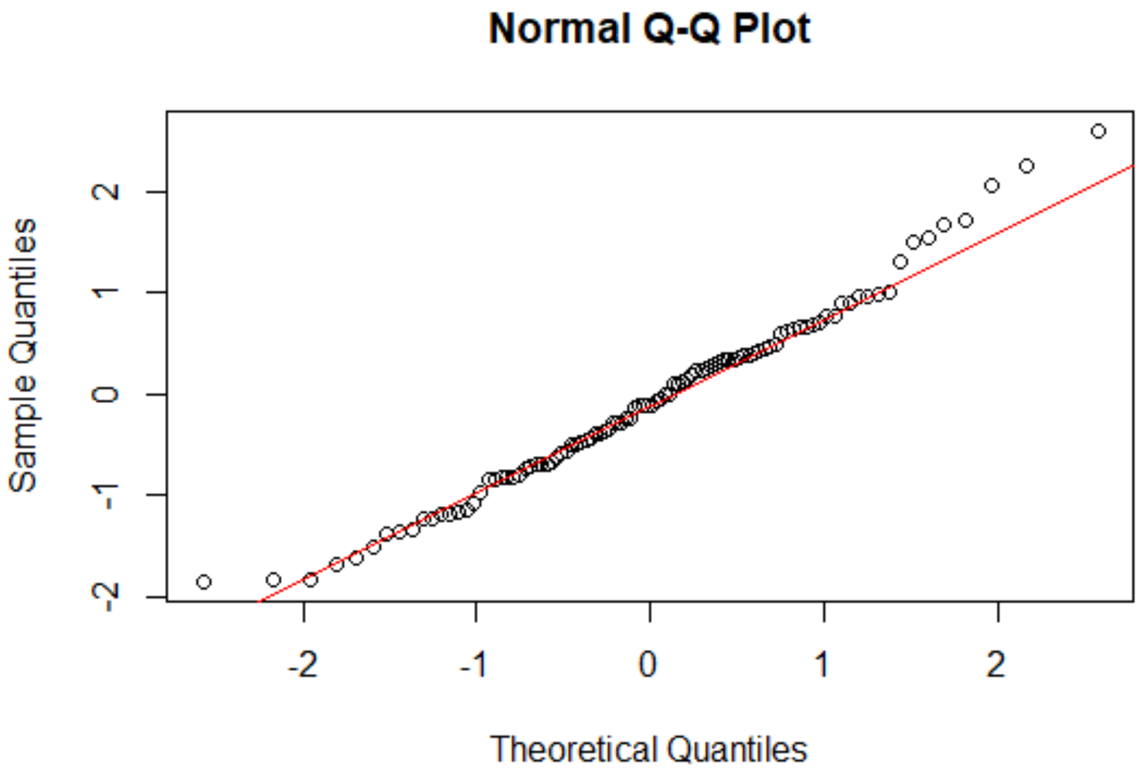
Step 1: Confirm Assumptions

- One sample compared to known value
- Testing for true unknown mean (μ)
- In this case we have unknown σ^2
- Do we have approximate normality??

1 Sample Inference – 1-Sample t and z tests

Confirming Normality

qqnorm(data)
qqline(data, col="red")



1 Sample Inference – 1-Sample t and z tests

Running the Test

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

1-Sample t-test

Assumptions:

-One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown μ (mean)
- Unknown σ^2

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

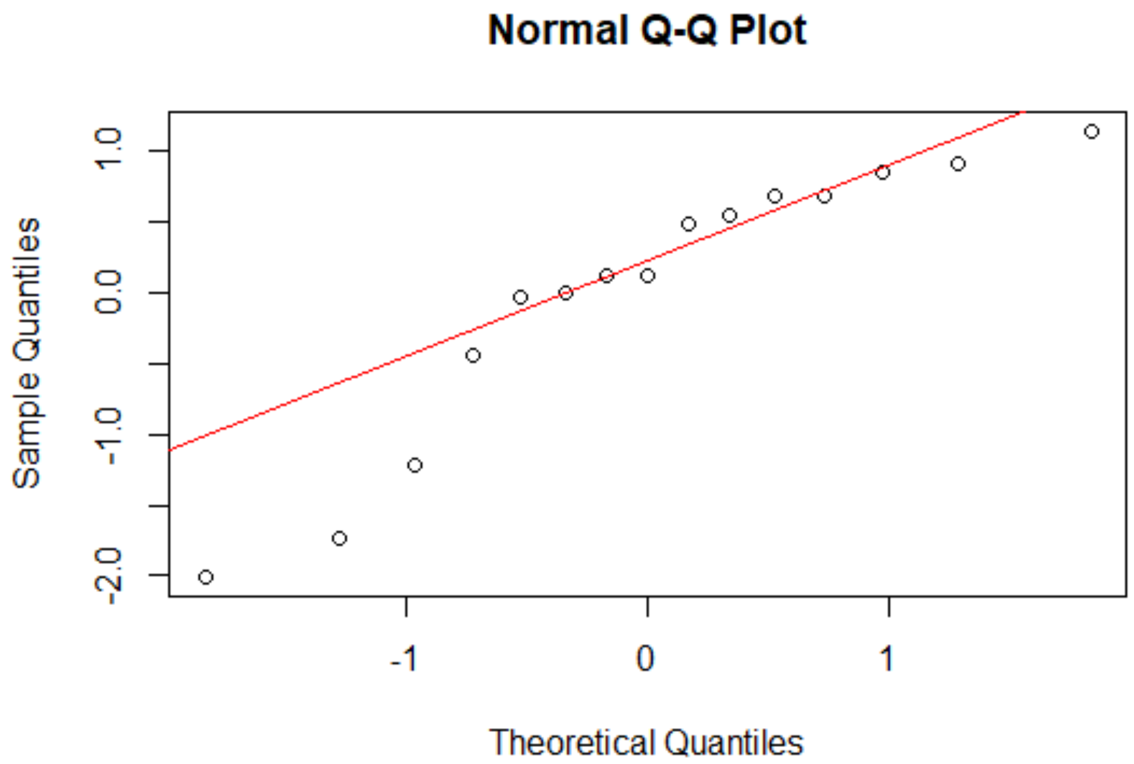
Test Stat: $t^* = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{s} = \frac{\sqrt{100}(23024 - 25000)}{6705} = -2.95$

Rejection Rule (in R):

$$
\begin{array}{ll}
1 - pt(t^*, n-1) & if \quad \mu > \mu_0 \\
pt(t^*, n-1) & if \quad \mu < \mu_0 \\
2(1 - pt(|t^*|, 0, 1) & if \quad \mu \neq \mu_0
\end{array}
$$

1 Sample Inference – Nonparametric Approach

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.

```
> 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
```

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):

$1 - pnorm(z^*, 0, 1) \quad if \quad m > m_0$

$pnorm(z^*, 0, 1) \quad if \quad m < m_0$

$2(1 - pnorm(|z^*|, 0, 1)) \quad if \quad m \neq m_0$

2 Sample Inference – Welch-Satterthwaite T-test

Not To Worry, We Have R!

First, Need Data Like This:

response	group
15	1
22	2
36	2
43	1
27	1
35	2

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

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: $t^* = \dfrac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\left(\frac{s_1^2}{n_1}\right) + \left(\frac{s_2^2}{n_2}\right)}}$

Rejection Rule (in R):

$$1 - pt(t^*, 0, 1) \qquad if \ \ \mu_1 > \mu_2$$

$$pt(t^*, 0, 1) \qquad if \ \ \mu_1 < \mu_2$$

$$2(1 - pt(|t^*|, v) \qquad if \ \ \mu_1 \neq \mu_2$$

Where $v = \dfrac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$ **= degrees of freedom**

2 Sample Inference – Pooled & Paired T-tests

Pooled T-test

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

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

Test Stat: $t^* = \dfrac{\overline{Y_1} - \overline{Y_2}}{\sqrt{\left(\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}\right)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$

Rejection Rule (in R):

$1 - pt(t^*, n_1 + n_2 - 2)$ $if \;\; \mu_1 > \mu_2$
$pt(t^*, n_1 + n_2 - 2)$ $if \;\; \mu_1 < \mu_2$
$2(1 - pt(|t^*|, n_1 + n_2 - 2))$ $if \;\; \mu_1 \neq \mu_2$

Paired t-test

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

$H_0: \mu_1 - \mu_2 = 0$ (interested in difference)
$H_A: \mu_1 - \mu_2 \neq 0$
$\mu_1 - \mu_2 > 0$
$\mu_1 - \mu_2 < 0$

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

Rejection Rule (in R):

$1 - pt(t^*, n-1)$ $if \;\; \mu_1 - \mu_2 > 0$
$pt(t^*, n-1)$ $if \;\; \mu_1 - \mu_2 < 0$
$2(1 - pt(|t^*|, 0, 1))$ $if \;\; \mu_1 - \mu_2 \neq 0$

2 Sample Inference – Pooled & Paired T-tests

R CODE

Welch=Satterthwaite (2-Sample T-Test)

```
> 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

```
> 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)

```
> 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

```
> 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
```

2 Sample Inference – Pooled & Paired T-tests

Testing For Equal Variance

H_0: variances equal
H_A: variances not equal

```
> 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.

2 Sample Inference – Business Application (Lead Scoring)

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

$n_2 = 290$

$\bar{x}_2 = \$78,547$

$s_2 = \$8,431$

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

2 Sample Inference – Business Application (Lead Scoring)

Thought Flow

1 or 2 Samples?	Step 1	2 Samples (Interested in Comparing Means)
Independent or Dependent Samples?	Step 2	Independent (Design Setup)
Unknown σ_1^2, σ_2^2?	Step 3	Yes, this is typical
Equal Variance?	Step 4	Equality of Variance Test
Both samples approx. Normal?	Step 5	Normal QQ Plots

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

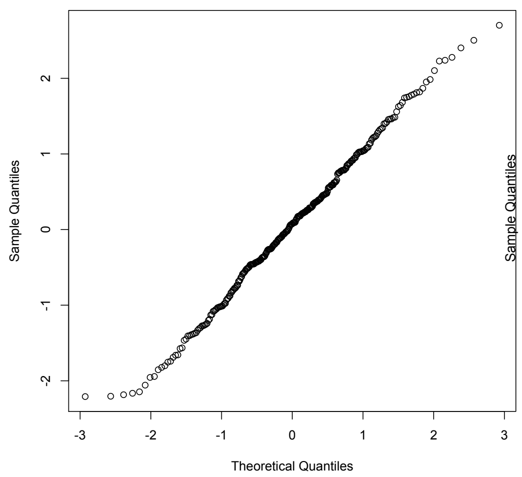
2 Sample Inference – Business Application (Lead Scoring)

```
> 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
```

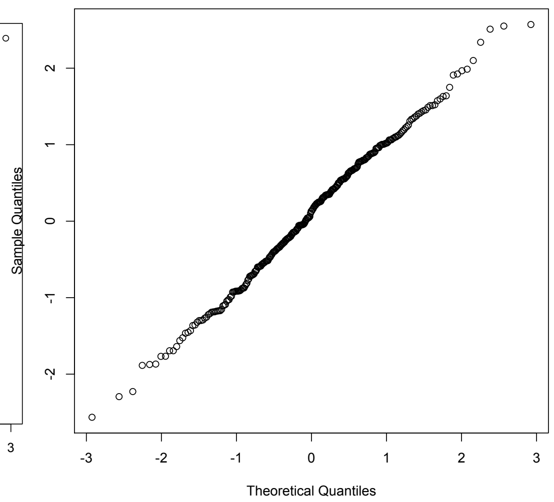
Conclusions

- -Variances are equal
- -Normality is satisfied

Normal Q-Q Plot - Think!



Normal Q-Q Plot - Superstructures



COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

2 Sample Inference – Business Application (Lead Scoring)

Paired t-test

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

1-Sample z-test

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

1-Sample t-test

***Assumptions*:**
- One sample compared to known value
- Data approx. normal (qqnorm in R)
- Unknown μ (mean)
- Unknown σ^2

Data Description

- 2 samples comparison of means
- Independent samples
- Unknown σ_1^2, σ_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 $\sigma_1^2 = \sigma_2^2$

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

2 Sample Inference – Business Application (Lead Scoring)

Pooled T-test

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

$H_A: \mu_1 > \mu_2 \ (\mu_1 - \mu_2 > 0)$

Where Group1=Think Architecture
Group 2=Susperstructures

```
> t.test(leads$Revenue~leads$LeadSource, alternative="greater", paired=FALSE, var.equal=TRUE)

	Two Sample t-test

data:  leads$Revenue by leads$LeadSource
t = -0.4097, df = 581, p-value = 0.6589
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 -0.1661357        Inf
sample estimates:
mean in group 1 mean in group 2 
     0.05335665      0.08644375 
```

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	*Sample 2*	*Sample 3*
x_1	y_1	z_1
x_2	y_2	z_2
x_3	y_3	z_3
$\vdots$	$\vdots$	$\vdots$
x_n	y_n	z_n

Analysis of Variance (ANOVA)

Resource:
https://onlinecourses.science.psu.edu/stat502/

Contact

Thank You!

Questions or Comments?