Spring 2021

SKKU Biostats and Big data

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

Lecture 19
Comparing Groups

CHOONG-WAN WOO | COCOAN lab | http://cocoanlab.github.io

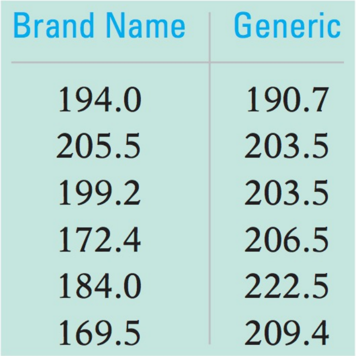
Review: Key Points

More about Tests and Intervals

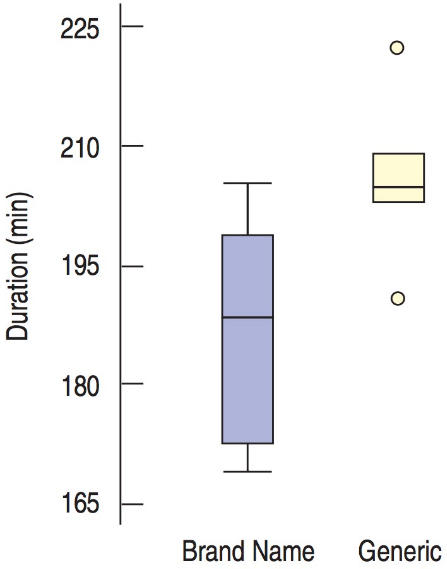
- **Type I error:** the null hypothesis is true, but we mistakenly reject it (false positive)
- **Type II error**: The null hypothesis is false, but we fail to reject it (false negative)
- Alpha: how small the P-value should be, P(Type I error)
- Beta: the probability of Type II error
- Power = 1 – beta
- Winner’s curse: increased bias in low powered studies
- Effect size: the distance between the null hypothesis value and the truth, but similar to signal-to-noise ratio

Standard deviation of a difference

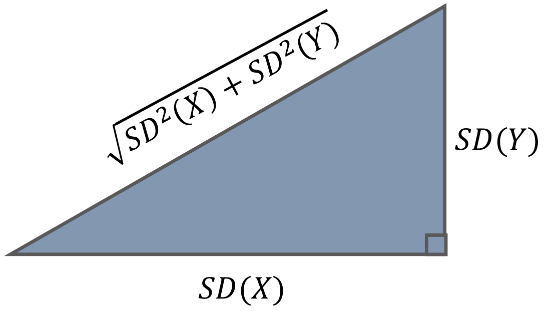
- Mean lifetime of brand-name vs. generic batteries:



Brand Name	Generic
194.0	190.7
205.5	203.5
199.2	203.5
172.4	206.5
184.0	222.5
169.5	209.4



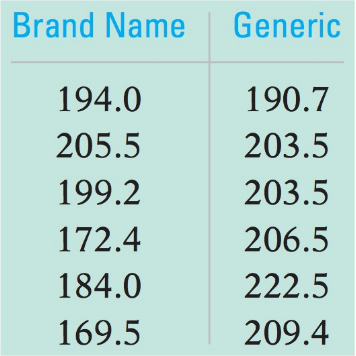
Pythagorean Theorem of Statistics



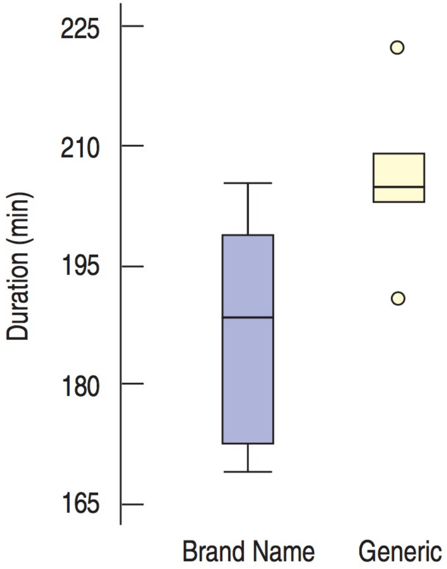
- We *observed* the difference between two groups.
- What’s the *true* difference for the general population?
- Pythagorean Theorem of Statistics: “*The variance of the sum or difference of two independent random variables is the sum of their variances.*”

Standard deviation of a difference

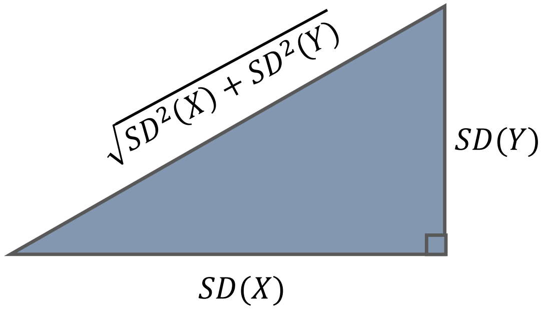
- Mean lifetime of brand-name vs. generic batteries:



Brand Name	Generic
194.0	190.7
205.5	203.5
199.2	203.5
172.4	206.5
184.0	222.5
169.5	209.4



Pythagorean Theorem of Statistics



- We *observed* the difference between two groups.
- What's the *true* difference for the general population?
- $Var(X - Y) = Var(X) + Var(Y)$

These works only for independent random variables.

- $SD(X - Y) = \sqrt{SD^2(X) + SD^2(Y)} = \sqrt{Var(X) + Var(Y)}$

Quiz 19-1

https://forms.gle/fo1ELjjP3tbtmBwX7

The standard deviation of the difference between two proportions

- $SD(\hat{p}_1) = \sqrt{\frac{p_1 q_1}{n_1}}$ and $SD(\hat{p}_2) = \sqrt{\frac{p_2 q_2}{n_2}}$
- $Var(\hat{p}_1 - \hat{p}_2) = (\sqrt{\frac{p_1 q_1}{n_1}})^2 + (\sqrt{\frac{p_2 q_2}{n_2}}\ )^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$
- $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$
- $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

Assumptions and Conditions for Comparing Proportions

- Independence Assumption
  - Randomization condition
  - The 10% condition

- Sample Size
  - Success/Failure Condition: at least 10 successes and 10 failures

- Independent Groups Assumption
  - Usually, this assumption is evident from the way the data were collected.
  - E.g., comparing husbands with their wives,

    or comparing subjects before vs. after some treatment

Quiz 19-2

https://forms.gle/6LBgL5wh3JoPebcC6

Confidence Interval for the Difference between two proportions

- Assuming the sampled values are independent, the samples are independent, and the sample sizes are large enough, the sampling distribution of $\hat{p}_1 - \hat{p}_2$ can be modeled by a Normal model with

  $\mu = p_1 - p_2$ and standard deviation $SD(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}$.

- Confidence interval: $(\hat{p}_1 - \hat{p}_2) \pm z^* \times SE(\hat{p}_1 - \hat{p}_2)$

- Standard error of the difference: $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1 \hat{q}_1}{n_1} + \frac{\hat{p}_2 \hat{q}_2}{n_2}}$

Two sample z-test: Testing for the differences between proportions

- Internet use before sleep:
  - 70.0% (205 of 293) of 19-29 years-old vs. 50.1% (235 of 469) of 30-45 years-old

- $H_0: p_1 - p_2 = 0$

- $SE(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2}}$
  - but *assuming* that the null hypothesis is true, $p_1 = p_2$, we need only single value for $\hat{p}$.
  - However, we have p_1 and p_2. We need to somehow combine these two proportions.
  - Pooling
    - combining the counts to get an overall proportion
    - $\hat{p}_{\text{pooled}} = \frac{Success_1 + Success_2}{n_1 + n_2}$, $SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_1} + \frac{\hat{p}_{\text{pooled}}\hat{q}_{\text{pooled}}}{n_2}}$
    - $\hat{p}_{\text{pooled}} = \frac{205+235}{293+469} = \frac{440}{762} = 0.5774$, $SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{0.5774 \times (1-0.5774)}{293} + \frac{0.5774 \times (1-0.5774)}{762}} = 0.0368$
    - $z = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{SE_{\text{pooled}}(\hat{p}_1 - \hat{p}_2)} = \frac{0.700 - 0.501}{0.0368} = 5.41$
    - $P = 2P(z > 5.41) \leq 0.0001$ (x 2 because it is a two-tailed test)

Confidence Interval for the Difference between two means

- $SD(\bar{y}_1 - \bar{y}_2) = \sqrt{Var(\bar{y}_1) + Var(\bar{y}_2)} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

- $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

- Two-sample *t*-interval: The sampling model is Student's *t* with adjusted degrees-of-freedom value

- $(\bar{y}_1 - \bar{y}_2) \pm ME$, where $ME = t_{df}^{*} \times SE(\bar{y}_1 - \bar{y}_2)$

$$\text{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

Two-Sample t-test

- $H_0: \mu_1 - \mu_2 = \Delta_0$
  - many times Δ_0 = 0

- $t = \frac{(\bar{y}_1 - \bar{y}_2) - \Delta_0}{SE(\bar{y}_1 - \bar{y}_2)}$

- $SE(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

$$\mathrm{df} = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{1}{n_1 - 1}\left(\frac{s_1^2}{n_1}\right)^2 + \frac{1}{n_2 - 1}\left(\frac{s_2^2}{n_2}\right)^2}$$

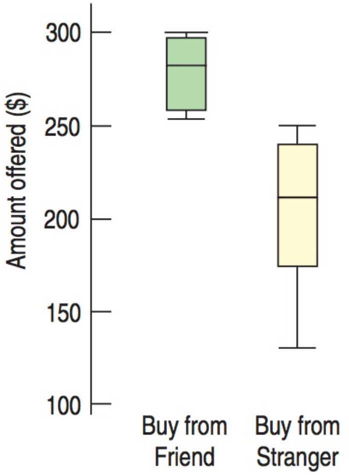
- When the conditions are met and the null hypothesis is true, the statistic can be closely modeled by a Student’s *t*-model with a number of degrees of freedom (adjusted). We use that model to obtain P-value.

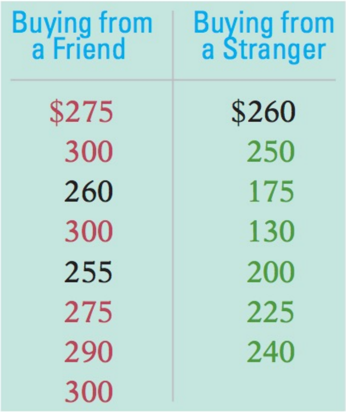
Distribution-free test: 1) Tukey's Quick Test

- John Tukey came up with a simpler alternative to the two-sample *t*-test
- Important numbers: 7, 10, and 13
- n_{high} = How many values in the high group are *higher* than all the values of the lower group?
- n_{low} = How many values in the low group are *lower* than all the values of the higher group?
- Count ties as ½
- If the total (n_{high} + n_{low}) > 7, similar to $\alpha = 0.05$, 10 and 13 gives us $\alpha = 0.01, 0.001$
- This quick test is used sometimes, but not accepted as the two-sample *t*-test.

n_{high} = 6.5 (1 tie: $260)
n_{low} = 6

12.5, thus P-value is between 0.01 and 0.001

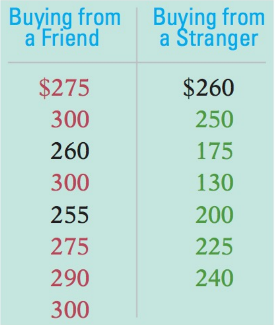




Buying from a Friend	Buying from a Stranger
$275	$260
300	250
260	175
300	130
255	200
275	225
290	240
300	

Distribution-free test: 2) Rank Sum test

- Wilcoxon rank sum (or Mann-Whitney) test
  - Less powerful than two-sample *t*-test, but it doesn't depend on the Nearly Normal Condition.
- Ranks the combined sample from the groups together from smallest to largest, assign 1 to N (= n_1+ n_2)
- If there are ties, use the average rank
- *W* is the rank sum of one group.
- Mean $\mu_W = \frac{n_1(N+1)}{2}$, variance $Var(W) = \frac{n_1 n_2(N+1)}{12}$, *z*-test with $z = \frac{W-\mu_W}{SD(W)}$

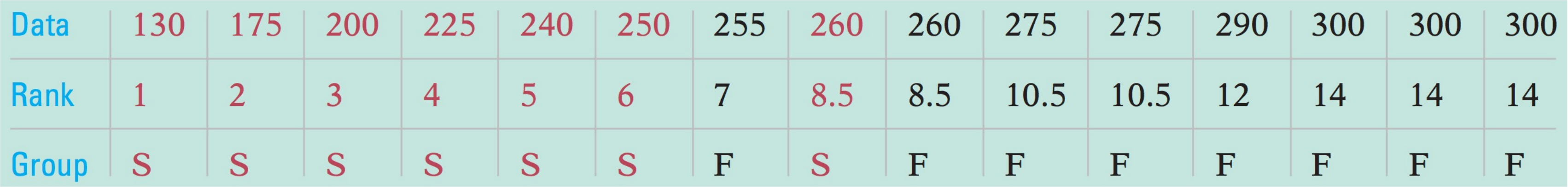


Buying from a Friend	Buying from a Stranger
\$275	\$260
300	250
260	175
300	130
255	200
275	225
290	240
300	

$$W = 7 + 8.5 + 10.5 + 10.5 + 12 + 14 + 14 + 14 = 90.5$$

$$\mu_W = \frac{8(15+1)}{2} = 64$$

$$SD(W) = \sqrt{Var(W)} = \sqrt{\frac{8 \times 7(15+1)}{12}} = 8.64, \text{so } z = \frac{90.5 - 64}{8.64} = 3.07$$



Data	130	175	200	225	240	250	255	260	260	275	275	290	300	300	300
Rank	1	2	3	4	5	6	7	8.5	8.5	10.5	10.5	12	14	14	14
Group	S	S	S	S	S	S	F	S	F	F	F	F	F	F	F

Quiz 19-3

https://forms.gle/Bk6i7pCmqgHN3rrP6

Pooled *t*-test

- This is simpler than two-sample *t*-test, but has a big assumption
  - "The variances of the two groups are the same."
  - **Advantages:**
    - This has a large degrees of freedom than two-sample t-test.
    - Simpler formula for degrees of freedom
  - **Disadvantages:**
    - The assumption of equal variances is a strong one, and is often not true, and difficult to check.

- $s^2_{\text{pooled}} = \frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{(n_1-1)+(n_2-1)}$
- $SE_{\text{pooled}}(\bar{y}_1 - \bar{y}_2) = \sqrt{\frac{s^2_{\text{pooled}}}{n_1} + \frac{s^2_{\text{pooled}}}{n_2}} = s_{\text{pooled}}\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$
- df = n_1+n_2-2

Key Points

Chapter 24: Comparing Groups

- $Var(X-Y)=Var(X)+Var(Y)$
- $SD(X-Y)=\sqrt{SD^2(X)+SD^2(Y)}=\sqrt{Var(X)+Var(Y)}$
- Confidence interval for the difference between two proportions: $(\hat{p}_1-\hat{p}_2)\pm z^*\times SE(\hat{p}_1-\hat{p}_2)$
  - $SE(\hat{p}_1-\hat{p}_2)=\sqrt{\frac{\hat{p}_1\hat{q}_1}{n_1}+\frac{\hat{p}_2\hat{q}_2}{n_2}}$
- Z-test for the difference between two proportions: $z=\frac{(\hat{p}_1-\hat{p}_2)-0}{SE_{\mathbf{pooled}}(\hat{p}_1-\hat{p}_2)}$
- Confidence interval for the difference between two means: $(\bar{y}_1-\bar{y}_2)\pm ME,$ where $ME=t^*_{df}\times SE(\bar{y}_1-\bar{y}_2)$
  - $SE(\bar{y}_1-\bar{y}_2)=\sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}}$
- Two-sample t-test, $t=\frac{(\bar{y}_1-\bar{y}_2)-\Delta_0}{SE(\bar{y}_1-\bar{y}_2)}$