STAT 500

Two-Sample Inference: Hypothesis Test

Scenario

- Randomized Experiment
  - Two treatments
  - Is there a difference in the mean value of the response variable between the two treatments?

- Observational Study
  - Two populations
  - One sample from each population
  - Is there a difference in the mean value of the variable between the two populations?

Notation

- Parameters
  - Population 1
    - $\boldsymbol{\mu_1}$ = mean value of variable in Population 1
    - $\boldsymbol{\sigma_1^2}$ = variance of variable in Population 1
    - $\boldsymbol{\sigma_1}$ = std. dev. of variable in Population 1
  - Population 2
    - $\boldsymbol{\mu_2}$ = mean value of variable in Population 2
    - $\boldsymbol{\sigma_2^2}$ = variance of variable in Population 2
    - $\boldsymbol{\sigma_2}$ = std. dev. of variable in Population 2

Notation

- Data
  - $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$

    value of variable for n_1 members from sample 1.
  - $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$

    value of variable for n_2 members from sample 2.

Notation

- Summary Statistics
  - Sample 1

    $\bar{Y}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Y_{1j}$

    $S_1^2 = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2 \qquad S_1 = \sqrt{\frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2}$

  - Sample 2

    $\bar{Y}_2 = \frac{1}{n_2} \sum_{j=1}^{n_2} Y_{2j}$

    $S_2^2 = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2 \qquad S_2 = \sqrt{\frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2}$

Research Question

- Do the two populations have the same mean value for the variable?

- Source of inference
  - Model-based inference.
  - Based on the distribution of test statistics.

Methods of Analysis

- Answer research question using
  - Visual displays
  - Statistical Summaries (means, std. devs., five number summaries)
  - Interval estimation: confidence interval for $\mu_1 - \mu_2$
  - Hypothesis Test: ($H_o : \mu_1 = \mu_2$)

Hypothesis Test

- $H_0 : \mu_1 = \mu_2$
- $H_A : \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$
- Assumptions
  - $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma_1^2)$
  - $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma_2^2)$
  - Y_{1j} and $Y_{2j'}$ are independent for all j and j'

Hypothesis Test

- Results
  - $\sum_{j=1}^{n_1} Y_{1j} \sim N(n_1\mu_1, n_1\sigma_1^2)$
  - $\sum_{j=1}^{n_2} Y_{2j} \sim N(n_2\mu_2, n_2\sigma_2^2)$
  - $\bar{Y}_1 \sim N(\mu_1, \sigma_1^2/n_1)$
  - $\bar{Y}_2 \sim N(\mu_2, \sigma_2^2/n_2)$
  - $\bar{Y}_1 - \bar{Y}_2 \sim N\left(\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$

Hypothesis Test

- Results

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

- In order to use above result for inference, we would need to know σ_1^2 and σ_2^2.

- σ_1^2 and σ_2^2 are population parameters and are generally unknown.

Estimation for Variances

$$S_1^2 = \frac{1}{n_1 - 1} \Sigma_{j=1}^{n_1} (Y_{1j} - \bar{Y}_1)^2 \text{ estimates } \mathrm{Var}(Y_{1j}) = \sigma_1^2.$$

$$S_2^2 = \frac{1}{n_2 - 1} \Sigma_{j=1}^{n_2} (Y_{2j} - \bar{Y}_2)^2 \text{ estimates } \mathrm{Var}(Y_{2j}) = \sigma_2^2.$$

Both estimators are unbiased estimators: $E(S_i^2) = \sigma_i^2$.

When $\sigma_1^2 \neq \sigma_2^2$ estimate $\mathrm{Var}(\bar{Y}_1 - \bar{Y}_2) = \dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$ as

$$\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$$

Estimation for Variances

- Additional Assumption

$$\sigma_1^2 = \sigma_2^2 = \sigma^2$$

- Estimate the unknown parameter σ^2 with S_p^2 (called the pooled sample variance).

$$\begin{aligned} S_p^2 &= \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2} \\ &= \frac{\Sigma_{j=1}^{n_1}(Y_{1j} - \bar{Y}_1)^2 + \Sigma_{j=1}^{n_2}(Y_{2j} - \bar{Y}_2)^2}{n_1 + n_2 - 2} \end{aligned}$$

- S_p^2 is an unbiased estimator of σ^2, i.e. $E(S_p^2) = \sigma^2$

Model-based Inference

- Assume each sample is a simple random sample from a population with a normal distribution, the samples are independent, and $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

- It follows that $$\frac{(n_1 + n_2 - 2)\, S_p^2}{\sigma^2} \sim \chi^2_{n_1+n_2-2}.$$

Hypothesis Test

- With equal variance assumption, we have

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim N(0, 1)$$

- Replacing σ with S_p gives

$$\frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- The distribution for the variable above is NOT $N(0, 1)$.

- What is the distribution?

Hypothesis Test

- Results
  - $\frac{(n_1+n_2-2)S_p^2}{\sigma^2} \sim \chi^2_{(n_1+n_2-2)}$
  - $\bar{Y}_1$ is independent of S_1^2
  - $\bar{Y}_2$ is independent of S_2^2
  - $\bar{Y}_1 - \bar{Y}_2$ is independent of S_p^2

Hypothesis Test

- Define two new variables

$$Z = \frac{\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \qquad W = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2}$$

- $Z \sim N(0, 1)$

- $W \sim \chi^2_{(n_1+n_2-2)}$

- Z is independent of W

Hypothesis Test

$$
\begin{aligned}
T &= \frac{Z}{\sqrt{W/(n_1+n_2-2)}} \\
&= \frac{\frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)}{\sigma\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}}{\sqrt{\left(\frac{(n_1+n_2-2)S_p^2}{\sigma^2}\right)/(n_1+n_2-2)}} \\
&= \frac{\bar{Y}_1-\bar{Y}_2-(\mu_1-\mu_2)}{S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}}}
\end{aligned}
$$

Hypothesis Test

$$T = \frac{Z}{\sqrt{W/(n_1 + n_2 - 2)}} \sim t_{n_1+n_2-2}$$

- t Distribution with $n_1 + n_2 + 2$ degrees of freedom
- Centered at zero
- Symmetric (mean and median are both zero)
- Bell-shaped
- More probability in the tails of distribution than $N(0,1)$
- As d.f. $\rightarrow \infty$, $t_{d.f.} \rightarrow N(0,1)$

Hypothesis Test

- $H_0 : \mu_1 = \mu_2$

- $H_A : \mu_1 \neq \mu_2$ or $\mu_1 < \mu_2$ or $\mu_1 > \mu_2$

- Assumptions
  - $Y_{11}, Y_{12}, \ldots, Y_{1n_1}$ are i.i.d. $N(\mu_1, \sigma_1^2)$
  - $Y_{21}, Y_{22}, \ldots, Y_{2n_2}$ are i.i.d. $N(\mu_2, \sigma_2^2)$
  - Y_{1j} and $Y_{2j'}$ are independent for all j and j'
  - $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Hypothesis Test

- Test Statistic

$$T = \frac{\bar{Y}_1 - \bar{Y}_2}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- If H_0 is true, $\mu_1 - \mu_2 = 0$, the test statistic T has a t distribution with $n_1 + n_2 - 2$ degrees of freedom.

- If H_0 is true, expect to observe T close to zero. Unlikely to observe a really large deviation from zero.

Hypothesis Test

- The p-value is the probability of obtaining the value of the test statistic or more extreme values (against H_0) if H_0 is true

- A small p-value means either:

  (i) H_0 is true and we were very unlucky

  or

  (ii) H_0 is false

Hypothesis Test

- $H_a : \mu_1 \neq \mu_2$

  p-value $= 2 * P(t_{n_1+n_2-2} > |T|)$

- $H_a : \mu_1 < \mu_2$

  p-value $= P(t_{n_1+n_2-2} < T)$

- $H_a : \mu_1 > \mu_2$

  p-value $= P(t_{n_1+n_2-2} > T)$

Hypothesis Test

- Scale of evidence

  $\boldsymbol{p > 0.10}$: unconvincing evidence of a difference

  $\mathbf{0.10} > \boldsymbol{p} > \mathbf{0.05}$: weak evidence

  $\mathbf{0.05} > \boldsymbol{p} > \mathbf{0.01}$: evidence of a difference

  $\mathbf{0.01} > \boldsymbol{p} > \mathbf{0.001}$: strong evidence

  $\boldsymbol{p} < \mathbf{0.001}$: very strong evidence

- The $\boldsymbol{p}$-value is NOT the probability that $\boldsymbol{H_0}$ is true

Lizard Infection Study

- Independent random samples of 15 lizards infected with a disease and 15 non-infected lizards
- Test null hypothesis that mean distance traveled in two minutes is the same for both populations
- t-test of $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 \neq \mu_2$

  $$t = \frac{-5.3733}{(7.4649)(\sqrt{\frac{1}{15}+\frac{1}{15}})} = -1.97 \text{ on } 28 \text{ d.f.}$$

  $$\Rightarrow p\text{-value} = 0.0586$$
- t-test of $H_0 : \mu_1 = \mu_2$ versus $H_0 : \mu_1 < \mu_2$

  $$t = \frac{-5.3733}{(7.4649)(\sqrt{\frac{1}{15}+\frac{1}{15}})} = -1.97 \text{ on } 28 \text{ d.f.}$$

  $$\Rightarrow p\text{-value} = 0.0293$$

```
/*  Part of the code posted as mlizards_ttest.sas that computes
    t-tests and related graphs  */

data set1;
  input lizard infection distance;
  datalines;
   1 1 16.4
   2 1 29.2
   3 1 37.1
   . .   .
   . .   .
   . .   .
  28 2 45.5
  29 2 24.5
  30 2 28.7
  run;

  proc format; value infection  1='yes'  2='no';
  run;
```

```
title1  'T-test for Mean Distance for Two Minute Runs';
  title2  'Sceloporis Occidentalis Lizards';
proc ttest data=set1;
  class infection;
  var distance;
  format infection infection.;
run;
```

T-test for Mean Distance for Two Minute Runs
Sceloporis Occidentalis Lizards

The TTEST Procedure

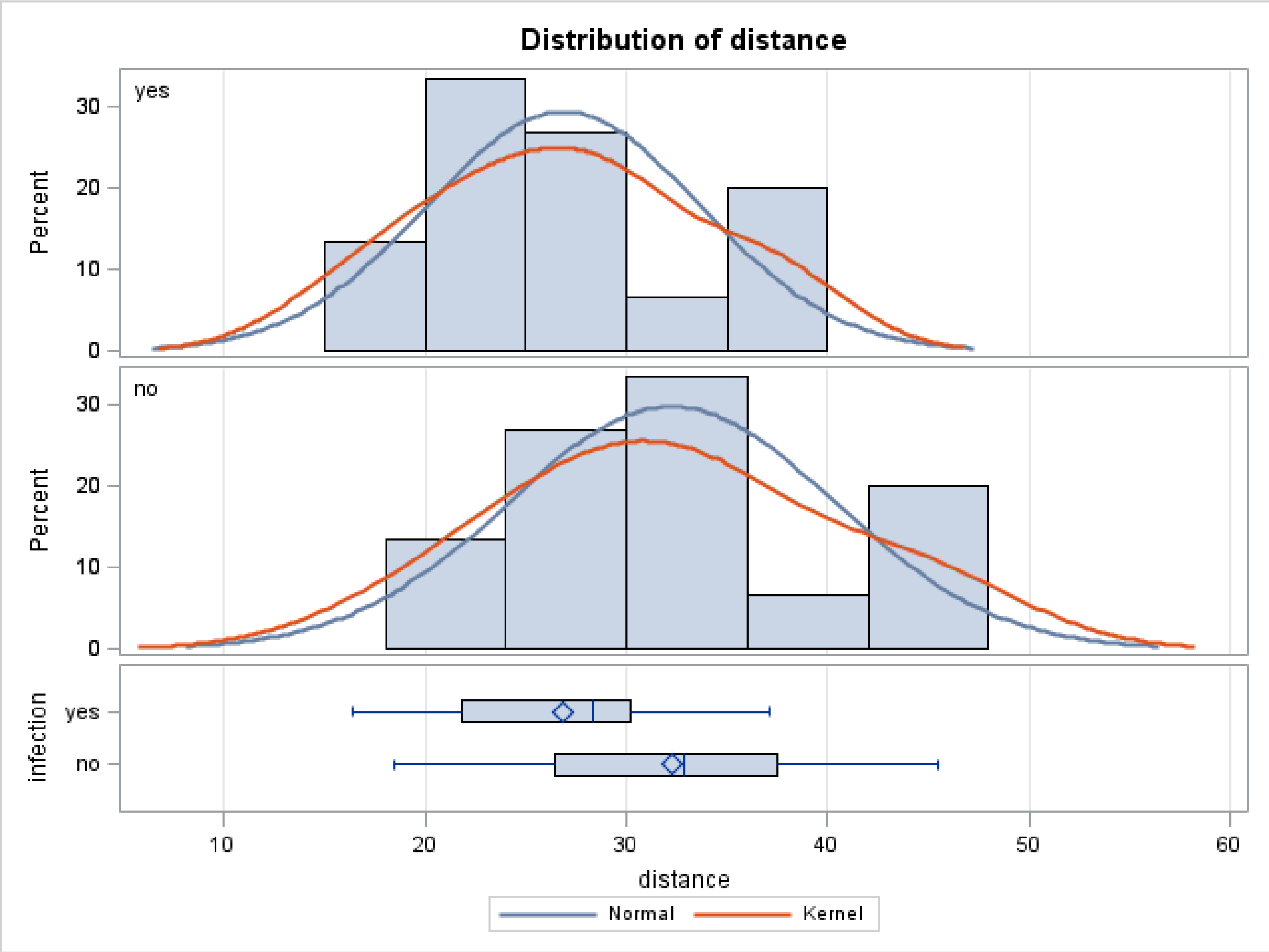
Variable: distance

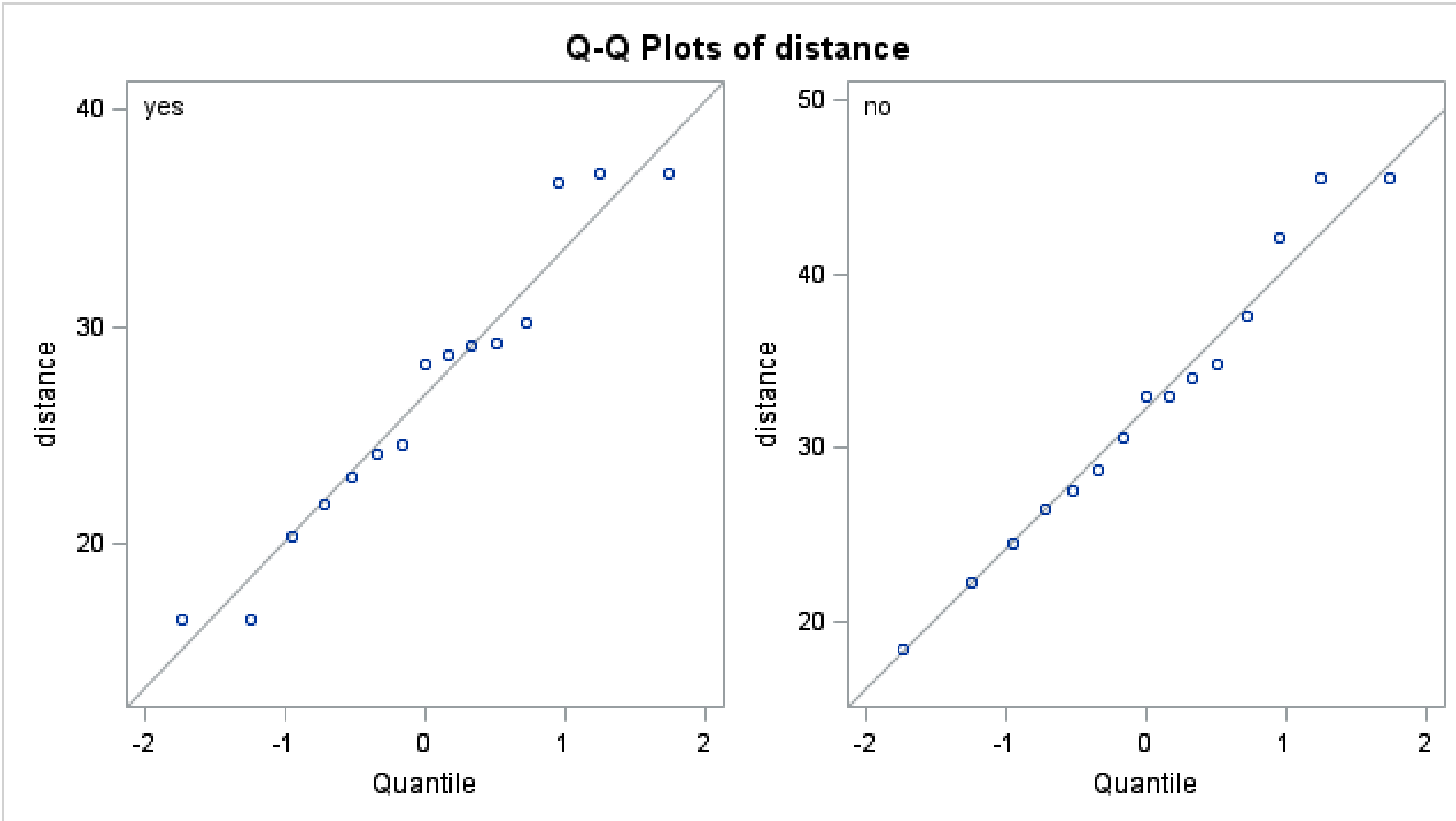
infection	N	Mean	Std Dev	Std Err	Minimum	Maximum
yes	15	26.8600	6.8096	1.7582	16.4000	37.1000
no	15	32.2333	8.0672	2.0829	18.4000	45.5000
Diff (1-2)		-5.3733	7.4649	2.7258		

infection	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
yes		26.8600	23.0889	30.6311	6.8096	4.9855	10.7395
no		32.2333	27.7659	36.7008	8.0672	5.9062	12.7228
Diff (1-2)	**Pooled**	-5.3733	-10.9569	0.2102	7.4649	5.9240	10.0960
Diff (1-2)	**Satterthwaite**	-5.3733	-10.9640	0.2173			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	28	-1.97	0.0586
Satterthwaite	Unequal	27.233	-1.97	0.0589

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	14	14	1.40	0.5343


Distribution of distance
yes
no
Percent
0
10
20
30
infection
10
20
30
40
50
60
distance
Normal
Kernel


Q-Q Plots of distance
yes
no
distance
Quantile
40
30
20
50
-2
-1
0
1
2