STAT 500

Two-Sample Inference
Model Assumptions and Diagnostics

Two-Sample Inference

- Model 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'
  - $\sigma_1^2 = \sigma_2^2 = \sigma^2$

Two-Sample Inference

- Summarize assumptions as:
  - Independence assumption
    - Each sample is i.i.d. (random sample)
    - Samples are independent
  - Homogeneous variance assumption
    - $\boldsymbol{\sigma_1^2 = \sigma_2^2 = \sigma^2}$
  - Normal Distribution assumption
    - Distribution of variable in each population is normal

Two-Sample Inference

- With assumptions, we have that

$$\boldsymbol{T = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2}}$$

- Result used to obtain
  - p-value for Hypothesis Test
  - Confidence Level for Confidence Interval

Model Assumptions

Those model assumptions may not exactly match with reality

- “No model is correct, but some are useful”
  George Box
- Expect some assumptions not to hold

Consequences of Violations of Assumptions

- If model assumptions are violated,
  - T will not have a $t_{n_1+n_2-2}$ distribution
  - Means that
    - p-value for hypothesis test will be wrong
    - Confidence Level for confidence interval will be wrong

Robustness of Two-Sample Inference

- How far off will true p-value and true Confidence Level be if model assumptions are violated?
  - Not far off - we can still use two-sample inference procedure.
  - Far off - we cannot use two-sample inference procedure.
- Research studies have established when violations of model assumptions will result in large differences between true p-values and confidence intervals compared to values obtained from two-sample inference procedure.

Model Diagnostics

Analysis of Assumptions

- Graphical and other evaluations of assumptions
- Understand robustness of methods
- If necessary, fix the problem
  - Transform or modify the data
  - Change the model
  - Use other statistical procedures

Model Diagnostics: Independence Assumption

- Study should be designed to achieve independent responses
- Independence may not hold if some sample members are related
  - students in same class
  - genetic relationships
  - soil samples taken close together
- Check by plotting observations versus relevant variables like time or location
- Look for possible clusters in which sample members may not respond independently

Model Diagnostics: Independence Assumption

- Two-Sample Inference Procedure is not robust to violating this assumption. Effects of correlated responses (random errors) include
  - Standard error formulas are incorrect
    $Var(\overline{Y}_1 - \overline{Y}_2) \neq \sigma^2(1/n_1 + 1/n_2)$
  - t procedures are in serious trouble
  - Confidence intervals will not have correct coverage probabilities
- Remedies - Use another statistical procedure
  - If clustering - reanalyze using appropriate methods
  - If time effects - use time-series models
  - If location effects - use spatial models

Model Diagnostics: Homogeneous Variances

- Graphical Methods
- Ratio of sample standard deviations
- Folded F-test for equality of variances
- Brown-Forsythe test

Graphical Methods

- Construct residual plots, histograms, or boxplots of values for each sample
- Look for
  - Outliers in each sample
  - Differences in IQR, Range
  - Differences in shape of sample distributions

Ratio of Sample Standard Deviations

$$\frac{max\{S_1,\ S_2\}}{min\{S_1,\ S_2\}}$$

- Between 1 and 2 - little impact
- Between 2 and 3 - potential impact
- Greater than 3 - likely impact

Folded F-test

- Test for equality of variances
- Reject $H_o : \sigma_1^2 = \sigma_2^2$ if

$$F_{max} = \frac{max\{S_1^2,\ S_2^2\}}{min\{S_1^2,\ S_2^2\}} \geq F_{(a,b),1-\alpha/2}$$

  where

  $a = n_1 - 1, b = n_2 - 1$ if $S_1^2 > S_2^2$

  $a = n_2 - 1, b = n_1 - 1$ if $S_2^2 > S_1^2$

- Very sensitive to normal distribution assumption
- Not recommended as the only check

Model Diagnostics
Homogeneous Variances

Output from TTEST procedure in SAS for the creative writing data

```
/*  Use the TTEST procedure to perform t-tests on
the creative writing dataset  */

title"Model Based t-tests";
proc ttest data=set1;
 class trt;
 var y;
 format trt trt.;
run;
```

trt	N	Mean	Std Dev	Std Err	Minimum	Maximum
intrinsic	24	19.8875	4.4418	0.9067	12.0000	29.7000
extrinsic	23	15.7391	5.2526	1.0952	5.0000	24.0000
Diff (1-2)		4.1484	4.8551	1.4167		

trt	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
intrinsic		19.8875	18.0119	21.7631	4.4418	3.4522	6.2308
extrinsic		15.7391	13.4677	18.0105	5.2526	4.0623	7.4343
Diff (1-2)	**Pooled**	4.1484	1.2950	7.0018	4.8551	4.0270	6.1152
Diff (1-2)	**Satterthwaite**	4.1484	1.2812	7.0156			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45	2.93	0.0053
Satterthwaite	Unequal	43.117	2.92	0.0056

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	22	23	1.40	0.4304

Brown-Forsythe Test

- Conduct a two sample t-test on the absolute deviations from the sample medians to assess homogeneous variability
- Available with the HOVTEST= option in the SAS GLM procedure
- *Sleuth*, Section 4.5.3 refers it to Levene's test
- Levene (1960) used absolute deviations from sample means

Brown-Forsythe Test

- Compute $Z_{1j} = |Y_{1j} - \text{median}_1|$ for $j = 1, ..., n_1$
  and $Z_{2j} = |Y_{2j} - \text{median}_2|$ for $j = 1, ..., n_2$

- Compute

$$\bar{Z}_1 = \frac{1}{n_1} \sum_{j=1}^{n_1} Z_{1j} \text{ and } S^2_{Z_1} = \frac{1}{n_1 - 1} \sum_{j=1}^{n_1} (Z_{1j} - \bar{Z}_1)^2$$

and

$$\bar{Z}_2 = \frac{1}{n_1} \sum_{j=1}^{n_2} Z_{2j} \text{ and } S^2_{Z_2} = \frac{1}{n_2 - 1} \sum_{j=1}^{n_2} (Z_{2j} - \bar{Z}_2)^2$$

Brown-Forsythe Test

- Compute $S^2_{Z \text{ pooled}} = \dfrac{(n_1 - 1)S^2_{Z_1} + (n_2 - 1)S^2_{Z_2}}{n_1 + n_2 - 2}$

- Reject $H_0 : \sigma_1^2 = \sigma_2^2$ if

$$\left| \frac{\bar{Z}_1 - \bar{Z}_2}{S_{Z \text{ pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right| > t_{(n_1+n_2-2),1-\alpha/2}$$

- Same as

$$\left| \frac{\bar{Z}_1 - \bar{Z}_2}{S_{Z \text{ pooled}} \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \right|^2 > F_{(1,n_1+n_2-2),1-\alpha}$$

```
/* Use the GLM procedure to perform the Brown-Forsythe
   test for homogeneous variances */

      proc glm data=set1 alpha=.05 ;
       class trt;
       model y = trt ;
       means trt / hovtest=bf;
       format trt trt.;
       run;
```

Brown and Forsythe's Test for Homogeneity of y Variance
ANOVA of Absolute Deviations from Group Medians

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
trt	1	3.8953	3.8953	0.36	0.5543
Error	45	493.7	10.9720		

Consequences of Unequal Variances

- Need to estimate quantity

$$Var(\overline{Y}_1 - \overline{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- Unbiased estimator is

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

- Two-Sample Inference procedure estimates this quantity as

$$S_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

Consequences of Unequal Variances

If $n_1 = n_2 = n$, two estimates are equal (df are different).

$$S_p^2 = \frac{(n-1)S_1^2 + (n-1)S_2^2}{2n-2} = \frac{S_1^2 + S_2^2}{2}$$

$$\begin{aligned} S_p^2\left(\frac{1}{n} + \frac{1}{n}\right) &= \left(\frac{S_1^2 + S_2^2}{2}\right)\left(\frac{2}{n}\right) \\ &= \frac{S_1^2}{n} + \frac{S_2^2}{n} \end{aligned}$$

Consequences of Unequal Variances

- Minor if sample sizes are equal, especially when df is large.
- Minor if the ratio of variances is within a factor of 4 (or 10).
- The worst case is when $\boldsymbol{n1 \ll n2}$ and $\boldsymbol{\sigma_1^2 > \sigma_2^2}$, i.e., the group with the smaller sample size has the larger variance.
  - Empirical type I error rate is not controlled at the nominal rate.
  - While controlling $\boldsymbol{\alpha = 5\%}$, the test may actually achieve 10% or 20% type I error rate.

Remedy: Approximate T-test

Use separate sample variances for the two samples. Then

$$\boldsymbol{T^* = \frac{\overline{Y}_1 - \overline{Y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}}$$

has an approximate $\boldsymbol{t}$-distribution with

$$\boldsymbol{\nu = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{1}{n_1-1}\left(\frac{S_1^2}{n_1}\right)^2 + \frac{1}{n_2-1}\left(\frac{S_2^2}{n_2}\right)^2}}$$

degrees of freedom. This is the Cochran-Satterthwaite approximation, and $\boldsymbol{\min(n_1 - 1, n_2 - 1) \leq \nu \leq n_1 + n_2 - 2}$

Remedy: Approximate T-test

- Welch test (*Sleuth* Section 4.3.2)
- Very similar results to two-sample inference when samples sizes are nearly equal
- Better performance with unequal sample sizes AND unequal variances

trt	N	Mean	Std Dev	Std Err	Minimum	Maximum
intrinsic	24	19.8875	4.4418	0.9067	12.0000	29.7000
extrinsic	23	15.7391	5.2526	1.0952	5.0000	24.0000
Diff (1-2)		4.1484	4.8551	1.4167		

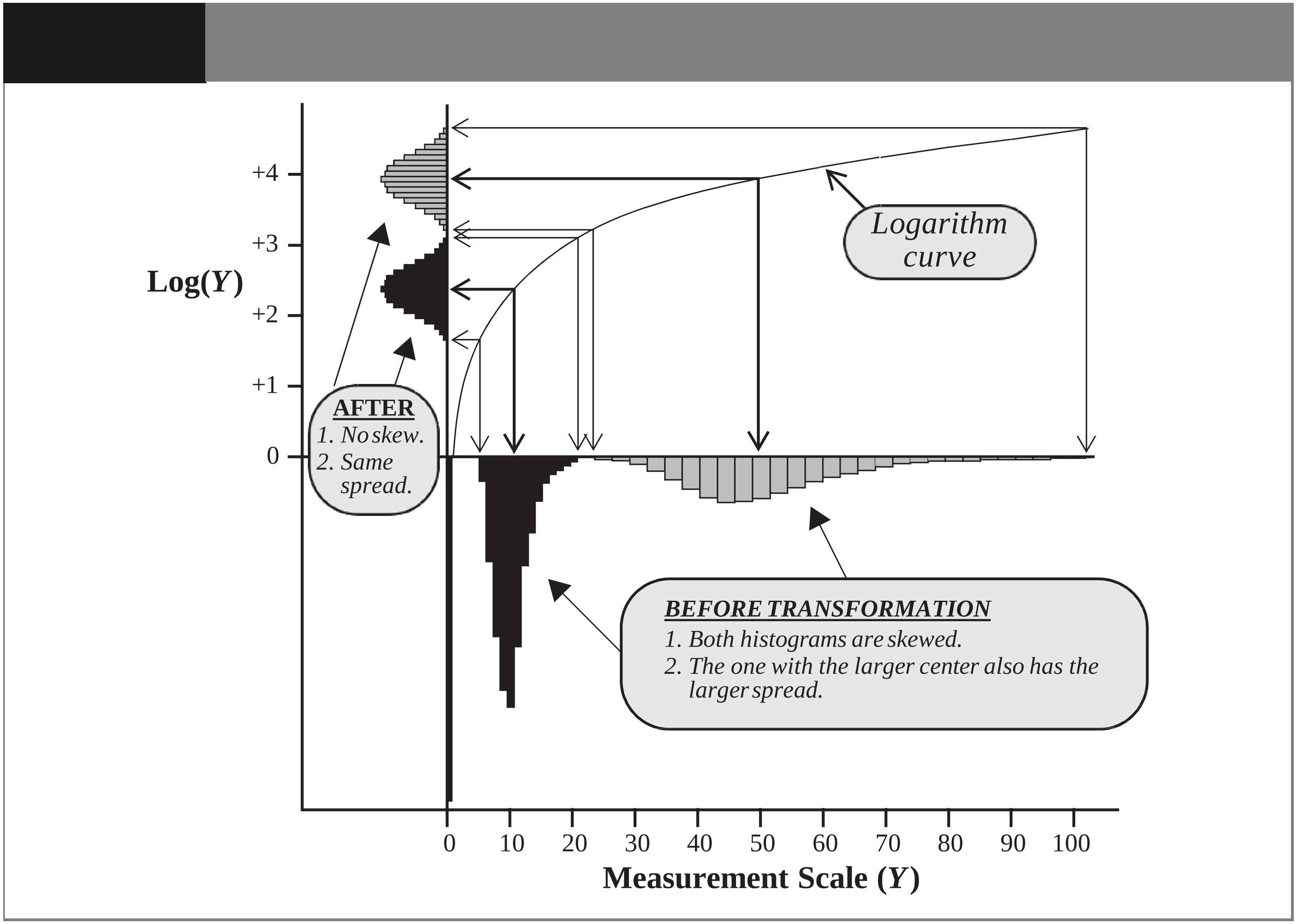
trt	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
intrinsic		19.8875	18.0119	21.7631	4.4418	3.4522	6.2308
extrinsic		15.7391	13.4677	18.0105	5.2526	4.0623	7.4343
Diff (1-2)	**Pooled**	4.1484	1.2950	7.0018	4.8551	4.0270	6.1152
Diff (1-2)	**Satterthwaite**	4.1484	1.2812	7.0156			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	45	2.93	0.0053
Satterthwaite	Unequal	43.117	2.92	0.0056

Equality of Variances				
Method	**Num DF**	**Den DF**	**F Value**	**Pr > F**
Folded F	22	23	1.40	0.4304

Variance Stabilizing Transformations

- Replace $\boldsymbol{Y_{ij}}$ with $\boldsymbol{X_{ij} = h(Y_{ij})}$
- Perform inference on $\boldsymbol{X_{ij}}$'s $\Rightarrow$ e.g., compare $\boldsymbol{\bar{X}_1}$ with $\boldsymbol{\bar{X}_2}$
- Back transform estimates to get conclusions on the $\boldsymbol{Y}$ scale
  - only approximate conclusions about population means on the $\boldsymbol{Y}$ scale
- How can $\boldsymbol{X_{ij} = h(Y_{ij})}$ affect heterogeneity?
  - Consider $\boldsymbol{h(Y) = log(Y)}$.



Log(Y)
+4
+3
+2
+1
0
Logarithm curve
AFTER
1. No skew.
2. Same spread.
BEFORE TRANSFORMATION
1. Both histograms are skewed.
2. The one with the larger center also has the larger spread.
0 10 20 30 40 50 60 70 80 90 100
Measurement Scale (Y)

Variance Stabilizing Transformations

- The logarithm function has a steep slope for small Y values, almost flat for large Y values.

- Small values are ‘stretched’ $\Rightarrow$ larger variance

- Large values are ‘shrunk’ $\Rightarrow$ smaller variance

Variance Stabilizing Transformations

- Choosing the transformation
  - Trial and error: try $\log(Y)$ or $\sqrt{Y}$
  - Rules of thumb:
    - data are all positive - use $\log(Y)$
    - data are proportions - use $arcsin(\sqrt{Y})$
    - data are counts - use $\sqrt{Y}$
  - transformation based on science (sqrt of area, cube root of volume)
  - Adjust for a variance-mean relationship; common for variance to increase with the mean

Variance Stabilizing Transformations

- If Var$(\boldsymbol{Y}) = \boldsymbol{g}(\boldsymbol{E}(\boldsymbol{Y}))$, then a variance stabilizing transformation can be obtained from

$$\boldsymbol{h}(\boldsymbol{y}) \propto \int \frac{\mathbf{1}}{\sqrt{\boldsymbol{g}(\boldsymbol{x})}}\, \boldsymbol{dx}$$

- Some Examples/Rules of Thumb
  - if var $\propto$ mean, then $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}$ and $\boldsymbol{h}(\boldsymbol{y}) = \sqrt{\boldsymbol{y}}$
  - if var $\propto$ mean$^{\mathbf{2}}$, then $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}^{\mathbf{2}}$ and $\boldsymbol{h}(\boldsymbol{y}) = \boldsymbol{log}(\boldsymbol{y})$
  - if var $\propto$ mean(1-mean), then $\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{x}(\mathbf{1} - \boldsymbol{x})$ and $\boldsymbol{h}(\boldsymbol{y}) = \sin^{-1}(\sqrt{\boldsymbol{y}})$

Variance Stabilizing Transformations

- Power transformation

  $$\boldsymbol{X} = \begin{cases} \boldsymbol{Y^{\lambda}} & (\boldsymbol{\lambda \neq 0}) \\ \boldsymbol{\ln Y} & (\boldsymbol{\lambda = 0}) \end{cases}$$

- estimate $\boldsymbol{\lambda}$ using maximum likelihood or use the variance - mean relationship

- Using variance-mean relationship
  - works for 2 groups or many groups (ANOVA)
  - Compute $\overline{\boldsymbol{Y}}$ and $\boldsymbol{S_Y}$ for each group
  - Regress log($\boldsymbol{S_Y}$) on log $\overline{\boldsymbol{Y}}$ and estimate the slope ($\boldsymbol{\beta}$)
  - Use transformation $\boldsymbol{Y^{\lambda}}$ with $\boldsymbol{\lambda = 1 - \beta}$

Power Transformation

- When does this work?
  - Population standard deviation is proportional to a power of the population mean:

    $$\boldsymbol{\sigma = \sqrt{Var(Y)} = \kappa\mu^{\beta}}$$

    or

    $$\boldsymbol{Var(Y) = \sigma^2 = \left[\kappa\mu^{\beta}\right]^2 = f(\mu)}$$

  - Use the delta method to obtain the transformation: $\boldsymbol{X = g(Y) = Y^{\lambda}}$

Power Transformation

Consider the Taylor series expansion

$$X = g(Y) \approx g(\mu) + (Y - \mu)g'(\mu)$$

Then an approximation for $Var(g(Y))$ is

$$Var(g(Y)) \approx [g'(\mu)]^2 Var(Y)$$

This is called the *delta method*

Power Transformation

- For $X = g(Y) = Y^{\lambda}$ we have

$$\frac{dX}{dY} = g'(Y) = \lambda Y^{\lambda - 1}$$

- From the delta method

$$\begin{aligned} Var(X) &\approx \left(\lambda\mu^{(\lambda-1)}\right)^2 \times \left(\kappa\mu^{\beta}\right)^2 \\ &= \kappa^2\,\lambda^2\,\mu^{2(\lambda-1+\beta)} \end{aligned}$$

- when $\lambda = 1 - \beta$ then $\lambda - 1 + \beta = 0$

  and $Var(X) \approx k^2\,\lambda^2$ is approximately constant

- Analyze the transformed data: e.g.,
  $X_{11} = log(Y_{11})$, $X_{12} = log(Y_{12})$, ..., $X_{2,n_2} = log(Y_{2,n_2})$

Power Transformations

- What if $\beta = 0.12$?
  Usually round to a "reasonable" value, i.e.,
  use $\beta \approx 0$ and $\lambda = 1$.

- Caution: Some researchers estimate the slope from the regression of $\log(Var(Y))$ on $\log(\overline{Y})$. Then use the transformation $Z = Y^{\lambda}$ with $\lambda = 1 - \beta/2$.

- Issues / concerns
  - Interpretation of results can be more difficult (e.g., the expectation of log(Y) is not the logarithm of $E(Y)$)

Model Diagnostics: Normality Assumption

- Graphical Methods
- Numerical Summaries
- Tests for Normality

Graphical Methods

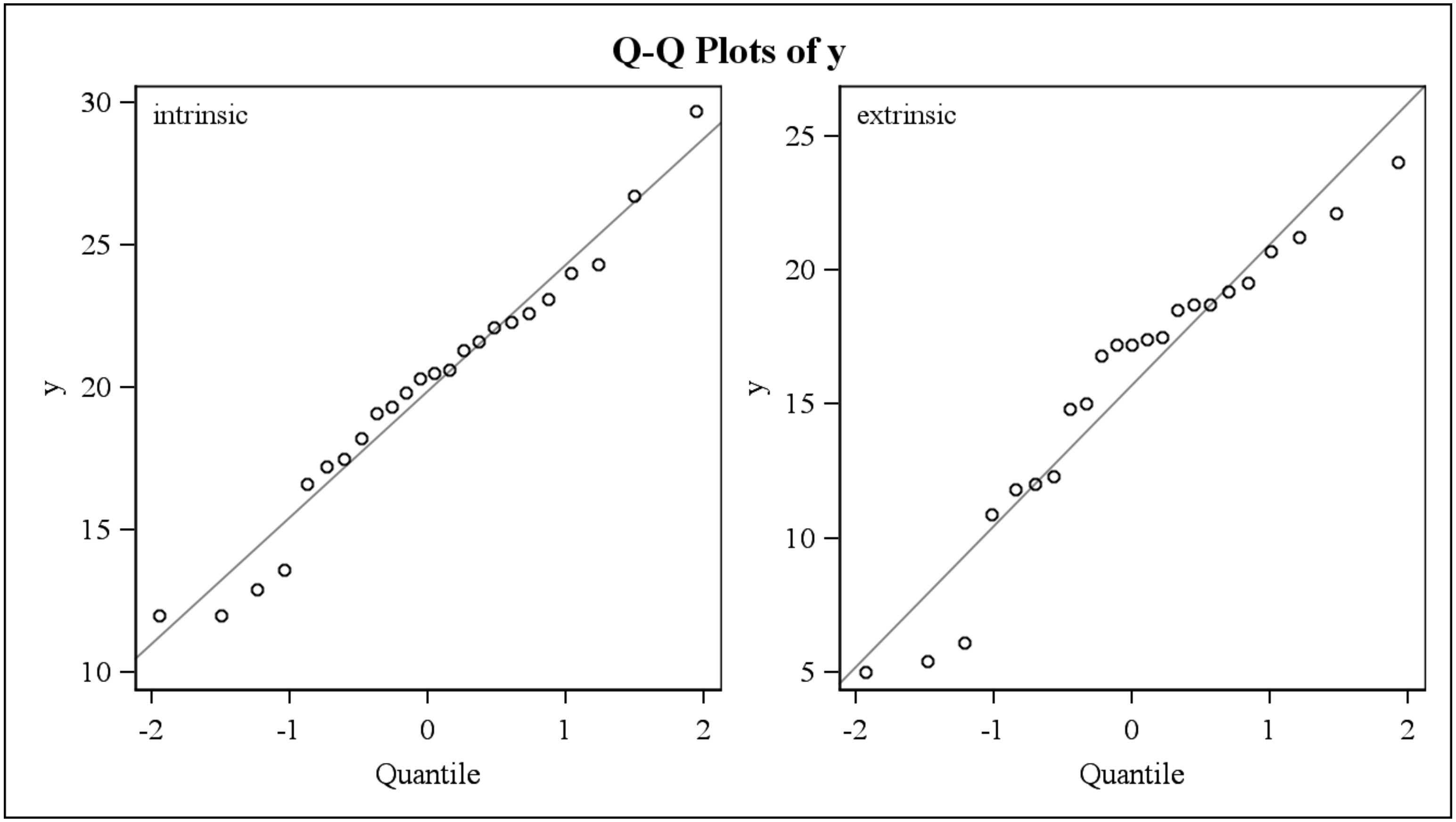
- Histogram of values from each sample
- Normal probability plot of values from each sample
  - Compare cumulative distribution function(CDF) of residuals to CDF for normal distribution
  - Most commonly done using quantiles: plot empirical quantiles (residuals) against expected quantiles from normal distribution

Normal Probability Plots

- Order observations (residuals) from smallest to largest (say $X_{(1)}, \ldots, X_{(n)}$)
- Compute expected quantiles ($q_{(1)}, \ldots, q_{(n)}$) from a standard normal distribution.
  - Expected quantiles can be calculated with tables.
  - General approximation: $q_i = \Phi^{-1}\left(\frac{i}{n+1}\right)$
  - Blom approximation $q_i = \Phi^{-1}\left(\frac{i-.375}{n+.25}\right)$
  - For $i = 5$, $n = 9$, $q_5 = \Phi^{-1}\left(\frac{5}{10}\right) = 0$
- Scatterplot of $X_{(i)}$ vs q_i should be close to a straight line with slope σ
- Curved patterns indicate non-normal distributions (or outliers)

The TTEST Procedure

Variable: y



Numerical Summaries

- For any normal distribution
  - skewness $= \boldsymbol{E(Y-\mu)^3/\sigma^3 = 0}$
  - Skewness measures the asymmetry

  - kurtosis $= \boldsymbol{E(Y-\mu)^4/\sigma^4 = 3}$
  - excess kurtosis = kurtosis - 3 (estimated by the UNIVARIATE procedure in SAS)
  - The sample kurtosis measures the heaviness of the tails of the data distribution.
  - positive value: long-tail; negative value: short-tail

Tests for Normality

- Many proposed tests for normality
- Tests based on empirical cdf's: Kolmogorov-Smirnov, Anderson-Darling, etc.
- Tests based on skewness or kurtosis
- Chi-square goodness-of-fit tests
- Tests based on normal probability plots: Shapiro-Wilk, correlation tests
- Normality is almost always rejected for large sample sizes.

Consequence of Non-Normality

- Large samples - very little (Central Limit Theorem)
- Small samples
  - Sample distributions have same shape and equal sample sizes - very little impact
  - Sample distributions have same shape and different sample sizes - potential impact if distributions are skewed
  - Sample distributions have different shapes - likely impact

Remedy: Non-normality

- Transformation (especially for skewness)
- Discussed earlier (under remedies for unequal variances)
- Detect and eliminate outliers
- Non-parametric tests

Model Diagnostics: Outliers

Outlier: one (or a few) very unusual observation(s).

- Always an issue if outliers are from a non-target population
- Goal: make inferences for the target population

  Data: from a mix of the target population and an outlier population
- Detect and eliminate outliers
- Reduce the effects of outliers by using "robust" procedures

Model Diagnostics: Outliers

- Analyze data with and without suspected outliers to see if inferences change

- Remove data only if one can argue that observations are from a different population. Remove any other observations from that different population.

- Acknowledge deletion of outliers in final report