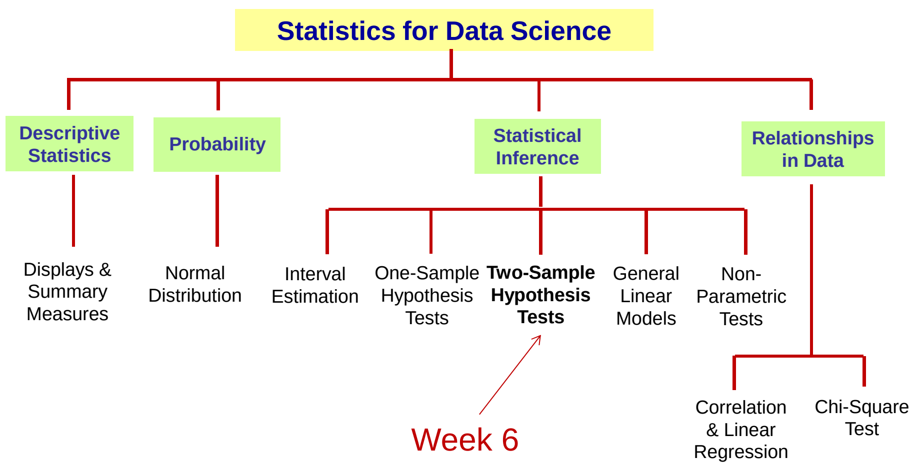
MATH 4044
Statistics for Data Science

Comparing Two Means

Course outline



Field, A & Miles, J, *Discovering Statistics Using SAS*, Chapter 9.

Topics to be covered

- Comparing two means:
  - Dependent t-test
  - Independent t-test
  - Confidence intervals and hypothesis tests
  - Checking assumptions

Looking at differences

- We are often interested in differences between groups of people or subjects.
- In experimental research, we often manipulate what happens to subjects so that we can make causal inferences.
  - The simplest experiment is one with only one independent variable (e.g. weigh loss) that is manipulated in only two ways (e.g. exercise, no exercise).
- This situation can be analysed with a *t*-test.

Rationale for the *t*-tests

- Two samples of data are collected and the sample means calculated.
  - These means might differ by either a little or a lot.
- If the samples come from the same population, then we expect their means to be roughly equal.
  - Although it is possible for their means to differ by chance alone, we would expect large differences between sample means to occur very infrequently.
- We use the standard error as a gauge of the variability between sample means.

Rationale for the *t*-tests

- If the difference between the samples we have collected is larger than what we would expect based on the standard error, then we can assume that:
  - There is no effect, sample means in our population fluctuate a lot and we have, by chance, collected two samples that are atypical of the population from which they came.
  - Or, the two samples come from different populations and the difference between samples is genuine rather than simply due to chance.
- If the null hypothesis is rejected, we gain confidence that the two sample means differ because of the different experimental manipulation imposed on each sample.

Comparing two means

- Dependent *t*-test
  - Compares two means based on related data, from 'matched' samples;
  - E.g., Data from the same people measured at different times.
- Independent *t*-test
  - Compares two means based on independent data;
  - E.g., data from different groups of people.

Dependent t-test

- One-sample *t*-test applied to differences
  - Identify how the differences will be calculated
- Assumptions
  - The two populations are dependent
  - The population of differences is Normal

- Set up the hypotheses and significance level
  - H_0: $\mu_d = 0$
  - H_1: $\mu_d \neq 0$ (or > 0 or < 0)
  - α = 0.05 (or 0.10 or 0.01)
- The test statistic for a dependent t-test is $t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}}$ with *n*-1 degrees of freedom.

Example: Car rentals

- A car rental company investigates its monthly data for each of its offices on variables such as revenue, number of rentals and average rental length.
- The monthly revenue data for its airport and city office in one Australian city for financial year from July 2007 to June 2008 was obtained.
- Is there a significant difference on average?

Example: Car rentals

Month	Airport	City	Difference
Jul-07	$ 283,591.00	$ 188,010.00	$ 95,581.00
Aug-07	$ 269,620.00	$ 197,874.00	$ 71,746.00
Sep-07	$ 312,220.00	$ 193,954.00	$ 118,266.00
Oct-07	$ 300,679.00	$ 210,545.00	$ 90,134.00
Nov-07	$ 217,889.00	$ 212,116.00	$ 5,773.00
Dec-07	$ 381,030.00	$ 277,022.00	$ 104,008.00
Jan-08	$ 232,288.00	$ 239,715.00	-$ 7,427.00
Feb-08	$ 186,285.00	$ 197,761.00	-$ 11,476.00
Mar-08	$ 230,672.00	$ 256,650.00	-$ 25,978.00
Apr-08	$ 248,172.00	$ 182,655.00	$ 65,517.00
May-08	$ 221,898.00	$ 146,602.00	$ 75,296.00
Jun-08	$ 257,073.00	$ 149,663.00	$ 107,410.00

The data are paired by month.

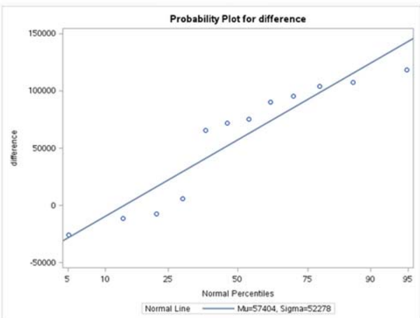
Example: Car rentals

- Is there a significant difference on average between the revenue at the airport office and the city office?
- Difference = Airport revenue – City revenue

- Hypotheses:
  - H_0: $\mu_d = 0$
  - H_1: $\mu_d \neq 0$
  - $\alpha = 0.05$
- Requirements:
  - As the sample size is small (n = 12 < 30), we need to test revenue differences for Normality.

Example: Car rentals

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.865318	Pr < W	0.0570
Kolmogorov-Smirnov	D	0.228329	Pr > D	0.0831
Cramer-von Mises	W-Sq	0.125342	Pr > W-Sq	0.0448
Anderson-Darling	A-Sq	0.712016	Pr > A-Sq	0.0468



The pattern in the P-P plot is not as straight as we would like, but Shapiro-Wilk (W = 0.87, *P*-value = 0.057) and Kolmogorov-Smirnov (D = 0.23, *P*-value = 0.083) tests suggest that the data can be assumed to be Normal, at least approximately.

We can proceed with a dependent *t*-test. ✓

Example: Car rentals

The TTEST Procedure

Difference: airport - city

N	Mean	Std Dev	Std Err	Minimum	Maximum
12	57404.2	52278.0	15091.4	-25978.0	118266

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
57404.2	24188.3	90620.0	52278.0	37033.5	88761.7

DF	t Value	Pr > |t|
11	3.80	0.0029

On average, the rental revenue at the airport office is significantly different from the rental revenue at the city office, $t(11) = 3.80$, P-value = 0.0029 < 0.05.

In fact, we are 95% confident that the revenue at the airport office is between \$24,188 and \$90,620 higher than the revenue at the city office.

Example: SAS code

```
data work.rental;
    input month $ airport city difference;
    datalines;
       Jul07 283591 188010 95581
       Aug07 269620 197874 71746
       Sep07 312220 193954 118266
       Oct07 300679 210545 90134
       Nov07 217889 212116 5773
       Dec07 381030 277022 104008
       Jan08 232288 239715 -7427
       Feb08 186285 197761 -11476
       Mar08 230672 256650 -25978
       Apr08 248172 182655 65517
       May08 221898 146602 75296
       Jun08 257073 149663 107410
    ; run;
```

First create a data file

For a dependent *t*-test in SAS, samples must be in separate columns

Example: SAS code

```
proc print data=work.rental;
run;
```

Check data file is correct

```
proc univariate data=work.rental normal;
    var difference;
    histogram;
    probplot / normal(mu=est sigma=est);
run;
```

Generate Normality tests

```
proc ttest data=work.rental;
    paired airport*city;
run;
```

Run a dependent *t*-test

Two independent samples

- The goal is to compare responses to two treatments or characteristics of two populations.
- A two-sample problem can arise from a randomized comparative experiment that randomly divides the subjects into two groups and exposes each group to different treatment.
  - There is no matching of the subjects in the two samples;
  - The two samples may be of different sizes.
- The responses in each group are independent of those in the other group.

Independent *t*-test (general)

- To test the null hypothesis that population means μ_1 and μ_2 are equal

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

- The two-sample *t*-statistic is given by $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

Standard error

- The alternative hypothesis is

$$H_1 : \mu_1 - \mu_2 \neq 0$$

- Requirements:
  - ☐ Two independent random samples;
  - ☐ Population distributions are Normal.

Independent *t*-test (general)

- The t-distribution is only an approximation for the distribution of the test statistic.
- Appropriate degrees of freedom are calculated by the Welch-Satterthwaite approximation:

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

- SAS and other statistical software will normally produce the numerical value for the degrees of freedom.
- If software is not available, a conservative approach is to use the smaller of $n_1 - 1$ and $n_2 - 1$ as the degrees of freedom.

Independent *t*-test (pooled)

- A more precise method if population variances can be assumed to be equal.
- It uses pooled standard deviation s_p, where

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

- The test statistic has $n_1 + n_2 - 2$ degrees of freedom:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{\bar{x}_1 - \bar{x}_2}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

Standard error

General or pooled t-test?

- When $n_1 = n_2$, pooled and unpooled standard errors are equal so the test statistic is the same for both procedures.
- As the pooled procedure is not exact unless population variances are equal, approximate degrees of freedom should be used.
- If sample sizes are equal or close, a pooled procedure with df = $n_1 + n_2 - 2$ is acceptable.
- When the sample sizes are very different, the pooled test can be misleading unless the sample deviations are similar.

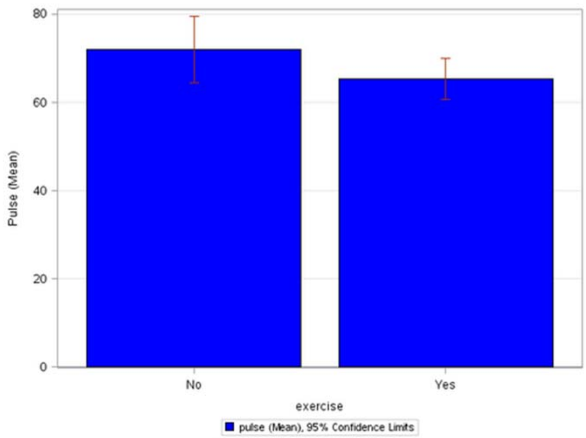
Example: Pulse rates

- It is believed that regular physical exercise leads to a lower resting pulse. Following are data for n = 20 randomly selected individuals on resting pulse rate and whether they exercise regularly or not.

Person	Pulse	Regularly exercises	Person	Pulse	Regularly exercises
1	72	No	11	62	No
2	62	Yes	12	84	No
3	72	Yes	13	76	No
4	84	No	14	60	Yes
5	60	Yes	15	52	Yes
6	63	Yes	16	60	No
7	66	No	17	64	Yes
8	72	No	18	80	Yes
9	75	Yes	19	68	Yes
10	64	Yes	20	64	Yes

Example: Pulse rates

- Is there a difference in pulse rate based on whether a person exercises regularly or not?



Tasks > Graph > Bar chart

Under *Options* go to *Statistics* and specify the type limits to be shown, confidence, standard error or standard deviation

Example: Pulse rates

- Let ‘No’ = 1 and ‘Yes’ = 2.
- Assumptions:
  - Two random and independent samples.
    - This requirement is satisfied.
  - Both populations should be Normal.
  - As the sample sizes are small (n_1 = 8 < 30 and n_2 = 12 < 30), we need to test both samples for Normality.

Example: Pulse rates

Exercise = ‘No’

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.924126	Pr < W	0.4642
Kolmogorov-Smirnov	D	0.15554	Pr > D	>0.1500
Cramer-von Mises	W-Sq	0.035066	Pr > W-Sq	>0.2500
Anderson-Darling	A-Sq	0.267799	Pr > A-Sq	>0.2500

Exercise = ‘Yes’

Tests for Normality				
Test	Statistic		p Value	
Shapiro-Wilk	W	0.947873	Pr < W	0.6061
Kolmogorov-Smirnov	D	0.237336	Pr > D	0.0612
Cramer-von Mises	W-Sq	0.079976	Pr > W-Sq	0.1952
Anderson-Darling	A-Sq	0.409296	Pr > A-Sq	>0.2500

P-values for all Normality tests are greater than 0.05, which suggests that both samples can be assumed to have come from Normal populations.

Therefore, we have all the requirements for an independent t-test. ✓

Example: Pulse rates

exercise	N	Mean	Std Dev	Std Err	Minimum	Maximum
No	8	72.0000	9.1339	3.2293	60.0000	84.0000
Yes	12	65.3333	7.4874	2.1614	52.0000	80.0000
Diff (1-2)		6.6667	8.1672	3.7278		

exercise	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
No		72.0000	64.3638	79.6362	9.1339	6.0391	18.5900
Yes		65.3333	60.5761	70.0906	7.4874	5.3040	12.7126
Diff (1-2)	Pooled	6.6667	-1.1652	14.4985	8.1672	6.1713	12.0779
Diff (1-2)	Satterthwaite	6.6667	-1.7274	15.0607			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	18	1.79	0.0906
Satterthwaite	Unequal	13.014	1.72	0.1099

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	7	11	1.49	0.5332

$H_0 : \mu_1 - \mu_2 = 0$

$H_1 : \mu_1 - \mu_2 \neq 0$

$\alpha = 0.05$

Using pooled t-test:
The test statistic is t = 1.79 (df = 18).
The corresponding P-value is 0.0906 > 0.05.
H_0 can't be rejected.

General or pooled *t*-test?

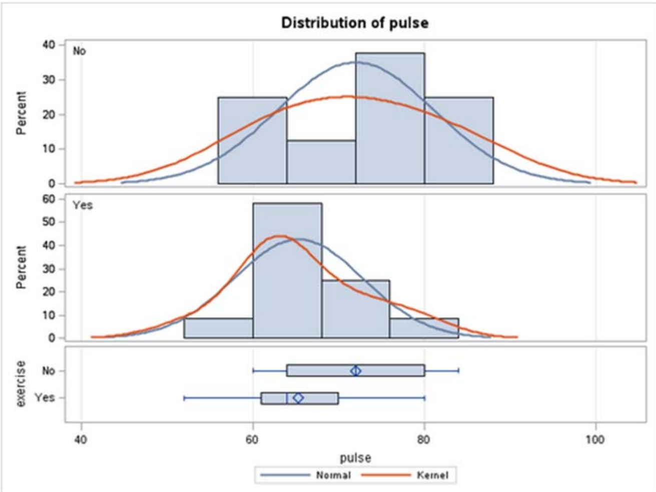
- The following advice is sometimes given:
  - Look at the table labelled Equality of Variances.
  - If the P-value is less than 0.05, then the assumption of homogeneity of variance has been broken.
    - For the t-test, use the P-value from the row labelled 'Satterthwaite'.
  - If the P-value is greater than 0.05, homogeneity of variance cannot be rejected.
    - For the t-test, use the P-value from the row labelled 'Pooled'.
- Caution:
  - Heterogeneity of variance is often accompanied by non-Normal distributions, and some tests of variances are not robust to their Normality assumption.

Example: Pulse rates

- On average, subjects who do not exercise regularly had higher pulse rates ($\bar{x}$ = 72, *SE* = 3.23) than subjects who do exercise regularly ($\bar{x}$ = 65.33, *SE* = 2.16).
- This difference was not significant at 5% level, $t(18)$ = 1.79, *P*-value = 0.0906 > 0.05.
- The 95% confidence interval for the difference in sample mean pulse rates is from -1.17 to 14.50.
  - As this confidence interval contains zero, this is another way to conclude that the sample difference between means is not statistically significant.

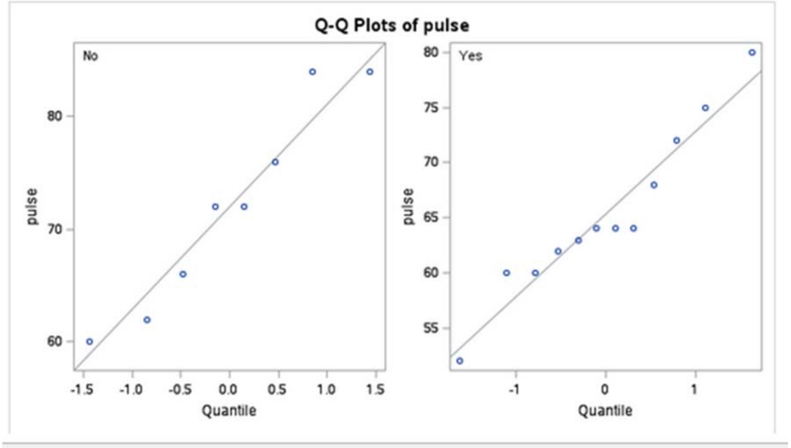
Example: Pulse rates

Graphical output of TTEST procedure



Example: Pulse rates

Graphical output of TTEST procedure



Example: SAS code

```
ods graphics on;

proc univariate data=work.pulse_ttest normal;
    var pulse;
    class exercise;
run;

proc ttest data=work.pulse_ttest;
    var pulse;
    class exercise;
run;

ods graphics off;
```

For an independent *t*-test in SAS, both samples need to be stored in a single column (`pulse`), with another column representing the group membership (`exercise`, ‘Yes’ or ‘No’).

Example: Comparing used car prices

- Websites of cars for sale such as www.carpoint.com.au include many variables in addition to price.
- Suppose we wish to compare the average price at dealers with that for private advertisers.
- For a particular model and year of manufacture the following sample statistics were obtained:

Advertiser	n	Mean	Std. Dev.
Dealer	23	$20,945	$3,004
Private	11	$17,934	$2,270

Example: Comparing used car prices

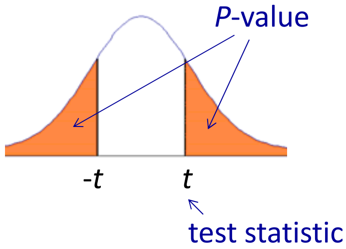
- Assumptions:
  - We have two random and independent samples.
  - Both samples come from Normal populations (needed since both samples are of size $n < 30$).
  - Population standard deviations are not equal.
- We will perform a two-sample t-test using the Welch-Satterthwaite approximation.
- Let ‘Dealer’ = 1 and ‘Private’ = 2. Then:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 \neq 0$$
$$\alpha = 0.05$$

Example: SAS code

```
/* Defining new temporary data file called ttest */
data work.ttest;

/* Defining variables to be stored in that file */
x1 = 20945; x2 = 17934; n1 = 23; n2 = 11;
sd1 = 3004; sd2 = 2270;

/* Calculating approximate degrees of freedom */
df = (sd1**2/n1 + sd2**2/n2 )**2/
     (1/(n1-1)*(sd1**2/n1)**2 + 1/(n2-1)*(sd2**2/n2)**2);

/* Calculating the t-test statistic */
t = (x1-x2)/sqrt(sd1**2/n1+sd2**2/n2);

/* Calculating P-value for a two-tailed test */
P_value = 2*(1-probt(abs(t),df));

/* Calculating 95% confidence limits */
CL_Left=(x1-x2) - TINV(.975,df)*sqrt(sd1**2/n1+sd2**2/n2);
CL_Right=(x1-x2) + TINV(.975,df)*sqrt(sd1**2/n1+sd2**2/n2);
run;

proc print data=work.ttest noobs;
run;
```


Example: P-value and confidence limits



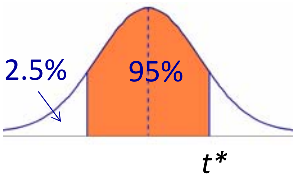
P-value $= P\,(t_{df} < -t\,) + P\,(t_{df} > t\,)$

$= 2 \times (1 - P\,(t_{df} < t\,))$

= `2*(1-probt(abs(t),df))`

Confidence interval:

Sample estimate $\pm$ Critical t-value t^* $\times$ Standard Error



$P\,(t_{df} < t^*\,) = 0.975 \quad \Rightarrow \quad t^*$

`TINV(.975,df)`

Example: Comparing used car prices

x1	x2	n1	n2	sd1	sd2	df	t	P_value	CL_Left	CL_Right
20945	17934	23	11	3004	2270	25.6024	3.24535	.003259672	1102.46	4919.54

- On average, dealers advertised higher prices ($\bar{x}$ = \$20,945, SE = 626.38) than private sellers ($\bar{x}$ = \$17,934, SE = 684.43) .
- The difference was statistically significant at 5% level, $t(25.6) = 3.25$, p-value = 0.003 < 0.05.
- We are 95% confident that the mean difference between dealer and private seller prices is between \$1,102.46 and \$4,919.54.

Dependent vs independent samples

- For same-subjects designs, the comparison of treatments is done within a subject thus eliminating differences between subjects from the comparison.
- For different-subjects designs, the random variation includes subject differences, which is likely to be larger than differences within the same subject.
  - Increased random variation in the outcome measures makes it more difficult to establish evidence of treatment differences.

t-test as a General Linear Model (GLM)

- Is there a relationship between a numerical variable (measurement of interest) and a categorical variable (group membership)?
- Recall from linear regression:

outcome = (model) + error

$$y_i = \underbrace{b_0 + b_1 x_i}_{\text{model}} + e_i$$

Simple linear regression

- In the pulse rates example:

Predictor

$$\underbrace{pulse_i}_{\text{Response}} = b_0 + b_1 \times \underbrace{group_i}_{\text{Dummy variable (1 = 'No', 0 = 'Yes')}} + e_i$$

Example: Pulse rates

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	213.33333	213.33333	3.20	0.0906
Error	18	1200.66667	66.70370		
Corrected Total	19	1414.00000			

The model is statistically significant at 10% level.

Root MSE	8.16723	R-Square	0.1509
Dependent Mean	68.00000	Adj R-Sq	0.1037
Coeff Var	12.01064		

Whether or not subjects exercise regularly explains 15% of variability in pulse rates.

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	65.33333	2.35768	27.71	<.0001
group	1	6.66667	3.72782	1.79	0.0906

The slope is statistically significant at 10% level.

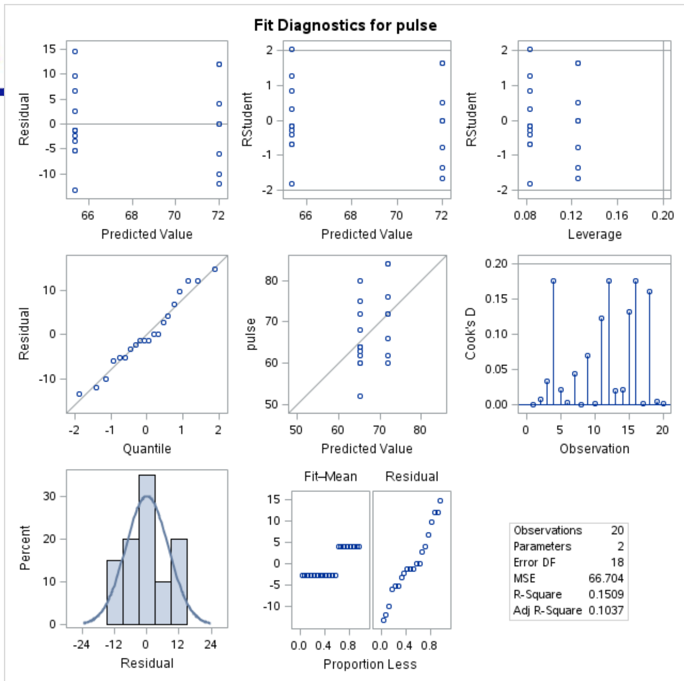
Example: Pulse rates

- 'Exercise' group is the baseline (*group* = 0).
- Intercept = mean of the baseline group:

$$\bar{x}_{\text{Exercise}} = b_0 + (b_1 \times 0)$$
$$b_0 = \bar{x}_{\text{Exercise}}$$
$$b_0 = 65.33$$

- Consider the 'No exercise' group (*group* = 1).
- b_1 = difference between means for the two groups:

$$\bar{x}_{\text{No exercise}} = b_0 + (b_1 \times 1)$$
$$\bar{x}_{\text{No exercise}} = \bar{x}_{\text{Exercise}} + b_1$$
$$\begin{aligned} b_1 &= \bar{x}_{\text{No exercise}} - \bar{x}_{\text{Exercise}} \\ &= 72.00 - 65.33 \\ &= 6.67 \end{aligned}$$



How do we interpret these diagnostic plots?

Example: SAS code

```
/* Defining dummy variables and creating a new temporary data
file called pulse_ttest_dummies */
data work.pulse_ttest_dummies;
    set work.pulse_ttest;

    if exercise='No' then
        group=1;
    else
        group=0;
run;

/* Simple linear regression using PROC REG with the dummy
variable ‘group’ as the only predictor */
proc reg data=work.pulse_ttest_dummies;
    model pulse=group;
    run;
quit;
```