# Single Sample and Paired Samples $t$ tests

### Dr. J. Kyle Roberts

Southern Methodist University
Simmons School of Education and Human Development
Department of Teaching and Learning

## Confidence Intervals for the Mean

- Suppose that we have a sample of IQ scores from 10 students.
- How might we know if the mean IQ of our 10 students is statistically significantly different from the population mean IQ of 100?

```
> iqdata <- c(100, 110, 95, 98, 120, 118, 143, 94,
+     82, 114)
> mean(iqdata)

[1] 107.4

> sd(iqdata)

[1] 17.35383
```
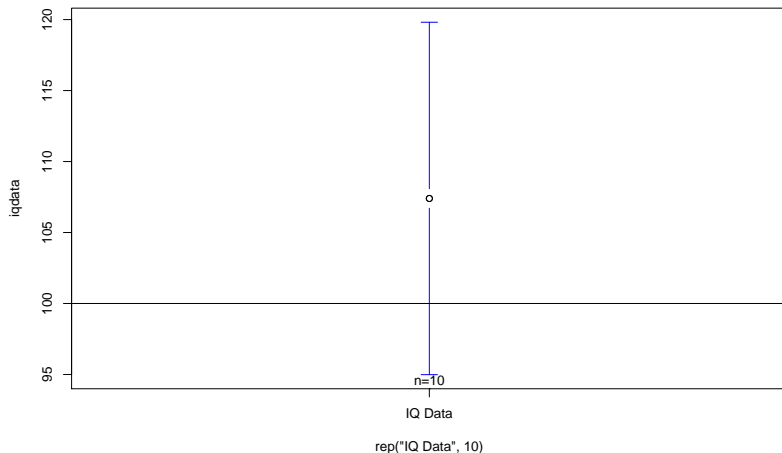
# IQ Data

We can then plot this data with a horizontal line representing the
population IQ score at 100.



rep("IQ Data", 10)

# Sampling from IQ Data

Suppose that we were to draw 25 random samples from the population of all IQ scores and draw 95% confidence intervals around each mean. It would stand to reason that 95% of the means would have confidence intervals that captured the population mean. For example:
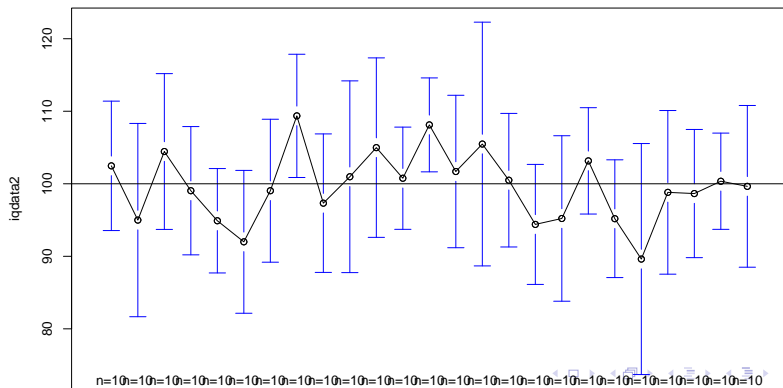
# More Study 1/Study 2 for Confidence Intervals

Suppose that we have two different datasets with the following
properties and a population mean of 20:

|      | Study 1 | Study 2 |
|------|---------|---------|
| mean | 10      | 15      |
| SD   | 10      | 10      |
| n    | 30      | 30      |

## Thought Question

Given the properties of the above two studies, which study would
be MORE likely to have confidence intervals which did not capture
the population mean?

# More Study 1/Study 2 for Confidence Intervals

Suppose that we have two different datasets with the following properties and a population mean of 20:

|      | Study 1 | Study 2 |
|------|---------|---------|
| mean | 10      | 10      |
| SD   | 5       | 10      |
| n    | 30      | 30      |

### Thought Question

Given the properties of the above two studies, which study would be MORE likely to have confidence intervals which did not capture the population mean?

# More Study 1/Study 2 for Confidence Intervals

Suppose that we have two different datasets with the following properties and a population mean of 20:

|      | Study 1 | Study 2 |
|------|---------|---------|
| mean | 10      | 10      |
| SD   | 10      | 10      |
| n    | 300     | 30      |

## Thought Question

Given the properties of the above two studies, which study would be MORE likely to have confidence intervals which did not capture the population mean?

# On Your Own

Try a dataset on your own. Consider the following data:

- $X = 7,8,7,6,4,2,3,8$

Complete the following tasks:

- Calculate confidence intervals (by hand).

- Determine if our data is statistically significantly different from a population mean of 10.

- Compute the confidence intervals in R by `ci(x)`.

# From CIs to $t$-calcs

- Recall from the Confidence Interval slides that we are looking to see if our sample mean captures the population mean given certain parameters.

- Rather than looking to see if our sample mean captures the population mean, we can compute the probability that our sample mean is different from the population mean.

## One-sample $t$ tests

- Occasionally we have a single sample of measurements and wish to evaluate a claim about the population mean, $\mu$, relative to some fixed, known value, $\mu_0$.

- If we assume that the sample, $x_1, x_2, \ldots, x_n$ is a random sample (i.e. the values of independent and identically distributed random variables) from a distribution with mean $\mu$ and variance $\sigma^2$ then, by the Central Limit Theorem, the random variable $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ tends to a $\mathcal{N}(0,1)$ distribution for large $n$.

- The quantity $\sigma/\sqrt{n}$ is called the emphstandard error of the mean (SEM). It describes the amount of variability we expect to see in $\bar{X}$.

- Unfortunately, we do not know $\sigma$.

- The empirical SEM is $s/\sqrt{n}$. That is, it uses the estimate $s$ of the (unknown) standard deviation $\sigma$.

## One-sample $t$ tests (continued)

- If the data, $x_1, x_2, \ldots, x_n$ are a random sample from a *normal* distribution, $\mathcal{N}(\mu, \sigma^2)$, then $\frac{\bar{X} - \mu}{s/\sqrt{n}}$ has a $T$ distribution with $n - 1$ *degrees of freedom*.

- The $T$ distribution is similar to the standard normal distribution but a bit more "spread out" or over-dispersed relative to the standard normal.

- In general we write the degrees of freedom as $df$. For a one-sample t-test $df = n - 1$ but for other tests we will have different formulas for $df$.

- The over-dispersion of the $T$ relative to the standard normal is greater for smaller values of $df$.

- For very large values of $df$, the $T$ is essentially the same as the standard normal.

## $p$-value of a statistical hypothesis test

- The $p$-value of a statistical hypothesis test is the probability of seeing the data that we did, or something even more unusual, if $H_0$ were true.
- However, $p$-values are somewhat akin to pregnancy tests. You either are or you are not.
- Some would argue that a very small $p$-value is strong evidence against $H_0$.
- In some disciplines predetermined "significance levels", such as 5% and 1%, are used. Thus a $p$-value less than 0.05 is "statistically significant".
- For the single-sample $t$ test, the $H_0 : \mu = \overline{X}$.

# One-sided versus two-sided alternatives

- Depending on the context of the data, we may be interested in an increase only ($H_a : \mu > \mu_0$), or a decrease only ($H_a : \mu < \mu_0$) or either type of a shift ($H_a : \mu \neq \mu_0$).

- An increase only or a decrease only are called one-sided alternatives.

- Either type of shift ($H_a : \mu \neq \mu_0$) is called a two-sided alternative.

- The intention to perform a one-sided test should be clear before the study is conducted and the data gathered. It should not be decided after examining the data.

- In many cases shifts in only one direction are desirable (we only wish to increase the yield, not decrease it; we only wish to decrease the proportion of defectives, not increase it) and the only sensible test is one-sided.

## Assumptions on the t-test

- For the results of `t.test` to be valid, we must assume that we have a random sample (i.e. independent observations with an identical distribution) **and** that the underlying distribution is a normal distribution.

- If the sample size, $n$, is large we can relax the requirement of a normal distribution to "sort-of like a normal".

- For small samples departures from normality become more important. However, when $n$ is small we cannot check reliably for departures from normality.

# Practical Example for IQ Data

Remember again our IQ Data
```
> iqdata <- c(100, 110, 95, 98, 120, 118, 143, 94,
+     82, 114)
> t.test(iqdata, mu = 100)

One Sample t-test
data:  iqdata
t = 1.3485, df = 9, p-value = 0.2105
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
  94.98581 119.81419
sample estimates:
mean of x
    107.4
```

- This means that our sample mean of 107.4 is not statistically significantly different from the population mean of 100 at the $\alpha = 0.05$ level on 9 $df$.

# Another IQ Data Example

This time consider the following dataset

```
> iqdata2 <- c(100, 110, 145, 115, 120, 118, 143,
+      94, 122, 114)
> t.test(iqdata2, mu = 100)

One Sample t-test
data:  iqdata2
t = 3.5325, df = 9, p-value = 0.006389
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 106.5089 129.6911
sample estimates:
mean of x
   118.1
```
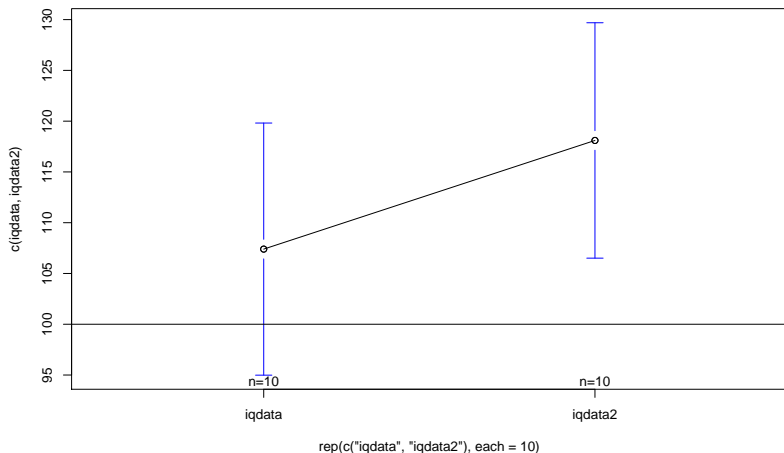
- This means that our sample mean of 118.1 *is* statistically significantly different from the population mean of 100 at the $\alpha = 0.05$ level on 9 $df$.

## Graphical Examples of `iqdata` and `iqdata2`

```
> plotmeans(c(iqdata, iqdata2) ~ rep(c("iqdata",
+     "iqdata2"), each = 10))
> abline(h = 100)
```

## Single Sample and Paired Samples $t$ test

- In the single sample $t$ test we are testing the null:

$$H_0 : \mu = \overline{X}$$

- It may be that we have a different data scenario where we have a set of scores gathered at two different time points.

- This is very common in education studies where some assessment instrument is given as a pre-test and post-test with an intervention between the two data gatherings.

- With a paired samples $t$ test, we can compute the statistical probability that the mean score at one time point is statistically significantly different from the mean score at another time point. Or:

$$H_0 : \overline{X}_{pre} = \overline{X}_{post}$$

# Paired Samples $t$ test Generalization

- Our original null hypothesis for the paired samples $t$ test is just a generalization of the single sample $t$ test null in that

$$H_0 : \overline{X}_{pre} = \overline{X}_{post}$$
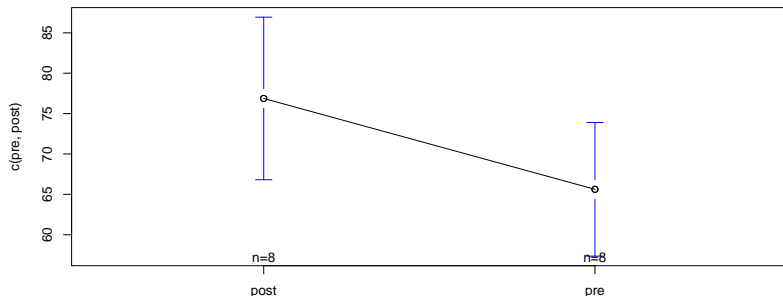$$\overline{X}_{pre} - \overline{X}_{post} = 0$$
$$\overline{X}_{diff} = 0$$

- This last form of the null hypothesis now takes on the same form as a single sample $t$ test where we are assuming the population mean difference score is 0.

# A Practical Example

```
> pre <- c(60, 70, 65, 68, 72, 49, 59, 82)
> post <- c(71, 79, 82, 78, 85, 60, 63, 97)
> c(mean(pre), sd(pre), mean(post), sd(post))

[1] 65.625000  9.898593 76.875000 12.040853

> plotmeans(c(pre, post) ~ rep(c("pre", "post"),
+     each = 8))
```
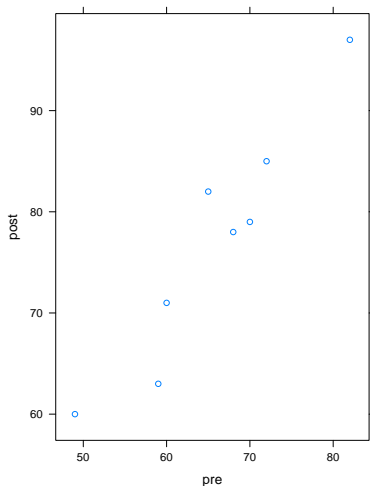
# Assumptions of the Paired Samples $t$ test

- Normality in the population.
- No interaction between pairs and treatment. This can happen when there is a ceiling effect in your data.
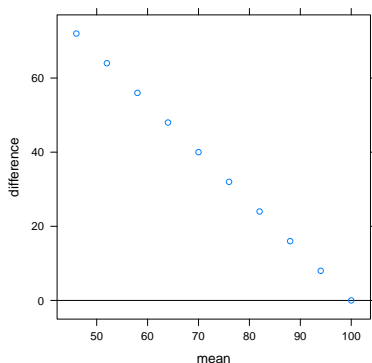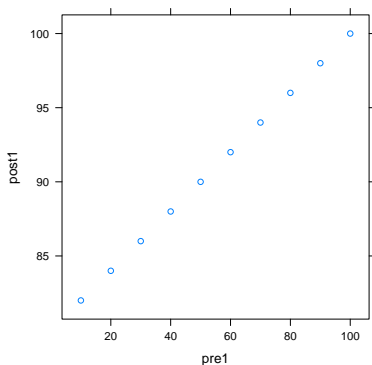- Outliers

## Assumptions of the Paired Samples $t$ test cont.

```
> print(p1 <- xyplot(post ~ pre), split = c(1, 1,
+     2, 1), more = TRUE)
> print(tmd(p1), split = c(2, 1, 2, 1))
```
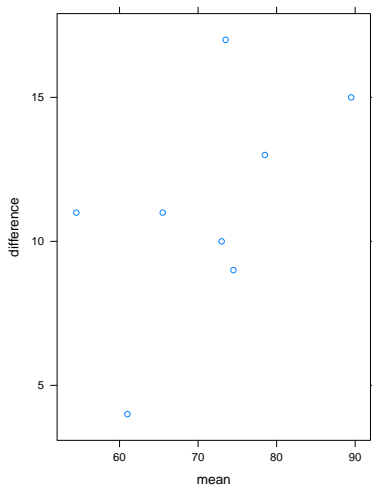
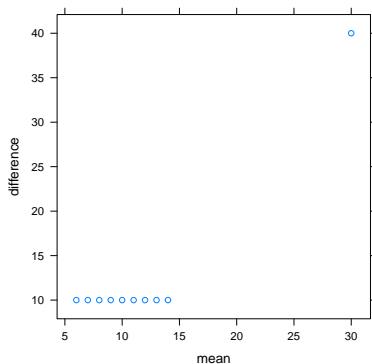## Assumptions with Interaction Data
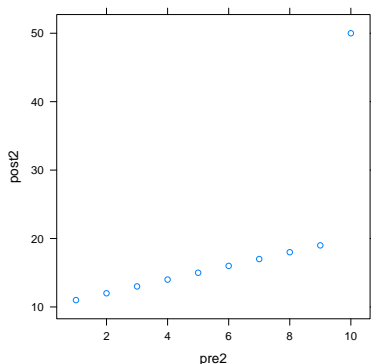
```
> pre1 <- seq(10, 100, by = 10)
> post1 <- seq(82, 100, length = 10)
> print(p2 <- xyplot(post1 ~ pre1), split = c(1,
+     1, 2, 1), more = TRUE)
> print(tmd(p2), split = c(2, 1, 2, 1))
```

## Assumptions with Outlier Data

```
> pre2 <- 1:10
> post2 <- c(11:19, 50)
> print(p3 <- xyplot(post2 ~ pre2), split = c(1,
+     1, 2, 1), more = TRUE)
> print(tmd(p3), split = c(2, 1, 2, 1))
```
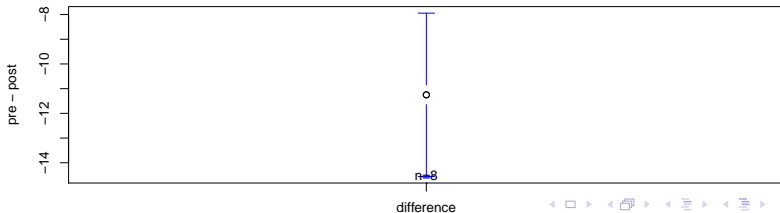
# Practical Example continued

```
> cbind(pre, post, diff = pre - post)

     pre post diff
[1,]  60   71  -11
[2,]  70   79   -9
[3,]  65   82  -17
[4,]  68   78  -10
[5,]  72   85  -13
[6,]  49   60  -11
[7,]  59   63   -4
[8,]  82   97  -15

> plotmeans(pre - post ~ rep("difference", 8))
```

# t.test in $R$

```
> t.test(pre, post, paired = T)

Paired t-test
data:  pre and post
t = -8.0452, df = 7, p-value = 8.792e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -14.55655  -7.94345
sample estimates:
mean of the differences
              -11.25

> t.test(pre - post, mu = 0)

One Sample t-test
data:  pre - post
t = -8.0452, df = 7, p-value = 8.792e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -14.55655  -7.94345
sample estimates:
mean of x
   -11.25
```