One sample t-test – test of the mean

Assumptions:

- Independent, random samples
- Approximately normal distribution
- (from intro class: σ is unknown, need to calculate and use s (sample standard deviation))

Hypotheses:

H_0: $\mu = \mu_0$ H_A: $\mu \neq \mu_0$ or use > or < in place of ≠

Most software, the default sign of the alternative hypothesis is ≠. In SAS (and other programs), the alternative can be changed.

$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ df=degrees of freedom = n - 1

Rejection Criteria:

Reject Ho if $pvalue \leq \alpha$

Assume that $\alpha = 0.05$ unless stated otherwise

General form of PROC TTEST:

```
PROC TTEST <options>;
   CLASS variable;
    VAR variable(s);
    PAIRED variables;
.
.
.
RUN;
```

Options for PROC TTEST statement:
DATA= SAS-dataset
ALPHA= specifies the significance level
H0= specifies the null value (μ_0)
SIDES= specifies one or two-tailed test
2, U, L
CI= requests confidence intervals for the standard deviation or the coefficient of variation
PLOTS produces statistical graphs (histograms and QQ plot)

2-sample t-tests

Independent samples:

- Pooled t-test – assumes that $\sigma_1^2 \approx \sigma_2^2$
- Unpooled t-test (Satterthwaite) – assumes $\sigma_1^2 \neq \sigma_2^2$

$H_0: \mu_1 = \mu_2$ $\quad$ $H_A: \mu_1 \neq \mu_2$ (can use > or <)

Or

$H_0: \mu_1 - \mu_2 = \Delta_0$ $\quad$ $H_A: \mu_1 - \mu_2 \neq \Delta_0$

Pooled t calculation:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

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

Unpooled t calculation:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

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

But…no thank you! ☺ A decent approximation is:

$df \approx \min(n_1 - 1, n_2 - 1)$ and SAS will calculate it for us.

Reject Ho if $pvalue \leq \alpha$

<u>Dependent samples (paired t-test)</u>:

Assumptions are the same as the independent samples tests but rather than independent random samples, the samples are dependent (still random) because each unit or subject is

measured twice – once before "treatment" and once after "treatment". The treatment does not have to be a literal treatment as it can be time (like GDP measured in countries over two time periods).

Hypotheses:

$H_0: \mu_D = \Delta_0$ $H_A: \mu_D \neq \Delta_0$ (can use > or <)
This test is done on the differences between the two measurements on each subject/unit. Once the differences are calculated, the test is equivalent to a one-sample test of the mean.

$d_i = x_i - y_j$ where x and y are the two measurements per subject.

$$\bar{d} = average\ difference = \frac{\sum d_i}{n}$$

$$s_d = st.dev.of\ differences = \sqrt{\frac{\sum d_i^2 - (\sum d_i)^2/n}{n-1}}$$

$t = \frac{\bar{d}-\Delta_0}{s_d/\sqrt{n}}$ and $df = n - 1$

Reject Ho if $pvalue \leq \alpha$

<u>Simple Linear regression (aka least-squares regression, etc.)</u>:

The main purpose of regression is to explore the relationship between 2 variables x and y, where x is the independent (explanatory) variable and y is the dependent (response) variable. Most often we calculate a regression equation ($y = mx + b$) to use in interpolation (estimation of y with an x that is in the known range of x values).

Population regression model:

$y = \beta_0 + \beta_1 x_i + \epsilon_i$ where:

β_0 is the intercept when x = 0
β_1 is the slope
ϵ_i is the random error (residual) term

Sample regression equation:

$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_i$

The "hats" mean estimated values. Note that the error term drops off. This is based on the assumptions of regression:

1. $E(\epsilon_i) = 0\ where\ \epsilon_i = y - \hat{y}$
   The residuals have a mean of 0.
2. $V(\epsilon_i) = \sigma_\epsilon^2$
   The variance of the residuals is homogeneous (constant for all values of $\hat{y}$).
3. $Cov(\epsilon_i, \epsilon_i') = 0$
   The residuals are independent (if the covariance between 2 variables is 0, then the variables are independent)
4. $\epsilon_i \sim N(0, \sigma_\epsilon^2)$
   The residuals should have an approximate normal distribution with mean 0 and variance σ_ϵ^2.

(the underlying assumption is that we do, in fact, have a linear relationship between x and y)

There is a lot more to regression than we get a chance to look at but this is a good start.

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

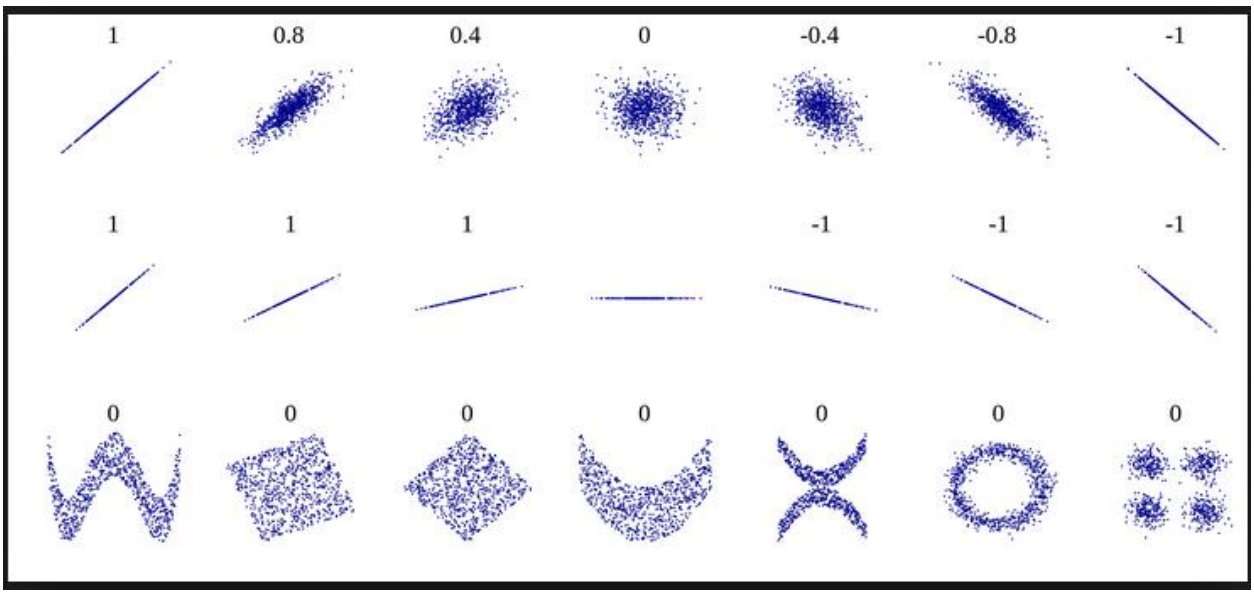
Correlation:

Population notation: ρ (rho)
Sample notation: r

While exploring the relationship between x and y, we will want to know how strong the relationship is. Correlation will explore the strength and direction (as in direction of the slope) of the ***linear*** relationship between x and y.

- Correlation is "unitless" as in it has no units of measurement associated with it
- $-1 \leq r \leq 1$ where the endpoints (-1,1) are perfect relationships and the closer it is to 0, the weaker the relationship is or a lack of relationship overall
- r makes no distinction between x and y
- units of measurement can be changed and it will not change the correlation

Some examples of different r values:



Ways to test if the relationship between x and y is significant (SAS does 2 out of the 3 ways in PROC REG):

- test the slope
- test the overall model
- test the correlation

1. Slope test:
   It is a t-test of the following hypotheses:
   $H_0: \beta_1 = 0 \quad H_A: \beta_1 \neq 0$ (can use < or > and the value can be something other than 0 if you are testing for a change in the slope rather than just significance of the slope)

   $t = \frac{\hat{\beta}_1 - 0}{se_{\hat{\beta}_1}}$ and $df = n - k$

   $where\ k = \#\ estimated\ parameters$
   $(in\ SLR, k = 2)$

2. Overall F test:
   It is a test of the following hypotheses about slopes:
   $H_0: \beta_1 = \beta_2 = \cdots = \beta_k = 0$
   $H_A: At\ least\ one\ \beta_k \neq 0$
   But, in SLR, there is only one slope so it is the same result as the t-test for the slope

   $$F = \frac{MSregression}{MSE}$$

3. Correlation test (not included in PROC REG):
   It is a test of the following hypotheses about the correlation:

   $H_0: \rho = 0 \quad H_A: \rho \neq 0$ (can use < or > and the value can be something other than 0 if you are testing for a change in the correlation rather than just significance of the correlation)

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

Experimental design for Complete Randomized Design (CRD):

The limitation of the 2-sample tests is what happens when you have more than 2 samples? You cannot just do several 2-sample tests because the Type I error rate will increase. Analysis of variance is for when you have more than two means to compare. However, its drawback is that it will not detect where the differences are, just that there are differences. The assumptions are the same as SLR.

Hypotheses:

$H_0: \mu_1 = \mu_2 = \cdots = \mu_t$
$H_A: at\ least\ one\ \mu_i\ differs$

Source	df	SS	MS	F
Treatment	t-1	SSTr	MSTr	MSTr/MSE
Error	N-t	SSE	MSE	n/a
Total	N-1	TSS	n/a	n/a

t = number of treatment groups (factor levels)
N = total number of observations in experiment
SSTr = Sum of Squares for treatment
MSTr = Mean square for treatment
SSE = Sum of squares for error (residuals)
MSE = Mean square error (aka residual variance σ_ϵ^2)

$$SSTr = \sum n_i\,(\bar{x}_{i.} - \bar{x}_{..})^2 \qquad SSE = \sum s_i^2\,(n_i - 1)$$

$$MSTr = \frac{SSTr}{t-1} \qquad MSE = \frac{SSE}{N-t}$$

The only thing ANOVA can do by itself is tell you that there is a difference in one of the group means and nothing else without a follow-up multiple comparison.