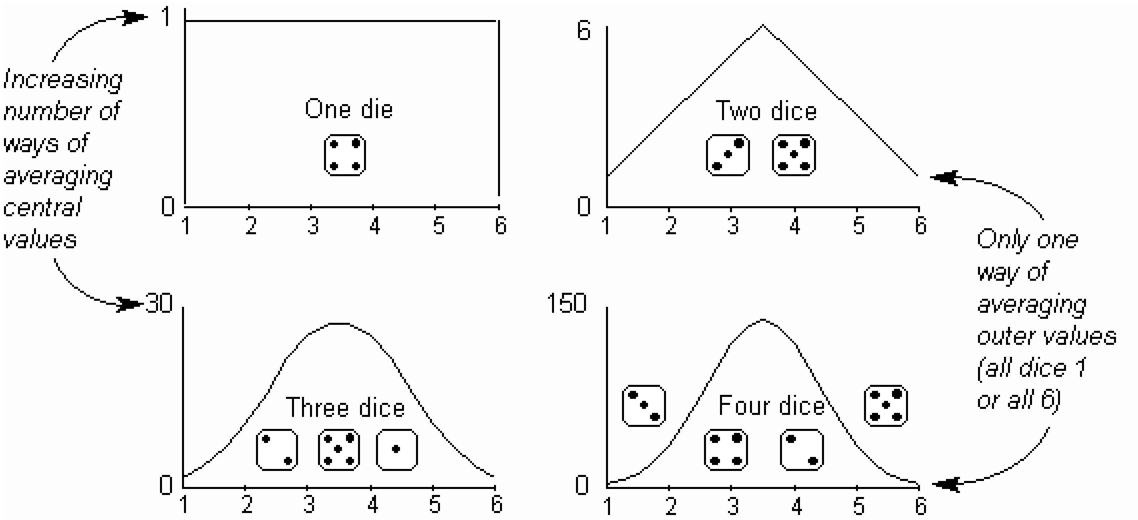
Unifications	of Techniques: See the Wholeness and Many-Fold ness
Population Continuous R.V.	μ, σ^2 Expected value and Variance both unknown Must be estimated with confidence. Homogenous population
Population Discrete R.V.	$\mu = \sum(xi\ pi) = E(x)$; $\sigma^2 = (\mathrm{E}x^2) - (Ex)^2 = \sum(xi^2\ pi) - [\sum(xi\ pi)]^2$; $\sigma = \sqrt{\sigma^2}$
Sample (s)	$\overline{x} = \frac{\sum x}{n}$; $s^2 = \frac{\sum(x-\overline{x})^2}{n-1} = \frac{SS}{n-1}, ss = \ Sum(x^2) - (Sum\ x)^2/$ n, C.V. = $\frac{s}{\overline{x}}$
Probability:	P(A given B)=P(A|B) = P(A and B)/ P (B) = $\frac{P(A\cap B)}{P(B)}$ P(A or B) = P(A∪B) = P(A) + P(B) - $P(A \cap B)$ A and B are independent if P(A|B) = P(A)

Binomial *Probability* *Distribution*	**x** success in **n** trial with $\boldsymbol{\pi}$ is probability success P(x) = $\frac{n!}{x!\,(n-x)!}\ \pi^x\ (1-\pi)^{n-x}$ μ =n π; σ^2 = n $\pi(1-\pi)$

Central limit theorem: Sampling distribution of mean tends to be normal density as the (fixed) sample size increases. μ=$\mu_{\overline{x}}$ and $\sigma_{\overline{x}}=\frac{\sigma}{\sqrt{n}}$ if σ *unknow, use* $\boldsymbol{s}$

CLT in Action:



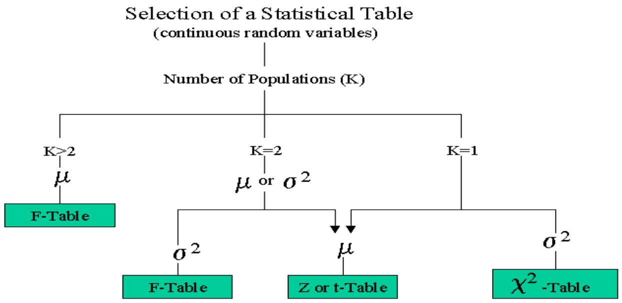
For the parent population the **expected value** is:

μ =$\sum(xi\ pi) = E(x)$ = 1(1/6) + 2(1/6) + 3(1/6) + 4(1/6) + 5(1/6) + 6(1/6) = 21/6 which the same for the other 3 sampling means distributions.

For the parent population the **variance** is:

$\sigma^2 = (\mathrm{E}x^2) - (Ex)^2 = \sum(xi^2\ pi) - (Ex)^2$ = 1(1/6) + 4(1/6) + 9(1/6) + 16(1/6) + 25(1/6) + 36(1/6) – $(21/6)^2$ = 4.04

The variance for the others is $\frac{\sigma^2}{n}$ = 4.04/2= 2.02, 4.04/3 = 1.37, 4.04/4 = 1.01, respectively, it get smaller as sample size increases.


Selection of a Statistical Table
(continuous random variables)
Number of Populations (K)
K>2
μ
F-Table
K=2
μ or σ²
K=1
σ²
F-Table
μ
Z or t-Table
σ²
χ²-Table

	For Population	For sampling distribution with sample size = n
Z-statistic	Z=$\frac{x-\mu}{\sigma}$ to make N(0,1)	$Z_{\bar{x}}$=$\frac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$ = $\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$ ~= $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ (in that $\mu_{\bar{x}} = \ \mu \ and \ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ and large n, invoke the CLT)
T-statistic	Normal	$T_{\bar{x}}$= $\frac{\bar{x}-\mu}{s/\sqrt{n}}$ with small n, but population is Normal

Notice the difference: The Z-score, Z-transformation = $\frac{x-\bar{x}}{s}$, is used to make data dimensionless, often for comparison purposes. For example price of the houses and their sized in Washington DC., and Tokyo.

Estimation with Confidence

	1 Population	2 Populations
Population mean μ	$\overline{\mathrm{x}} \pm z_{\frac{\alpha}{2}}\ \sigma_{\overline{x}} = \overline{x} \pm z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$ $\overline{x} \pm t_{\frac{\alpha}{2}}\ \sigma_{\overline{x}} = \overline{x} \pm t_{\frac{\alpha}{2}} \frac{s}{\sqrt{n}}$ (with v = n-1)	$\overline{x_1}$ - $\overline{x_2} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ $\overline{x_1}$ - $\overline{x_2} \pm t_{\frac{\alpha}{2}} \sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$ (With pooled estimate for S: s_p^2 =$\frac{(n_1-1)s_1^2+(n_2-1)s_2^2}{n_1+n_2-2}$) (v = $n_1 + n_2 - 2$)
Population proportion (probability) (p=$\frac{x}{n}$) Sample are sufficiently large	μ_p = π true population proportion of success. $\sigma_p = \sqrt{\frac{\pi\ (1-\pi)}{n}}$ P ± $z_{\frac{\alpha}{2}} \sqrt{\frac{p\ (1-p)}{n}}$	P_1 - $P_2 \pm z_{\frac{\alpha}{2}}\ \sigma_{(p1-p2)}$ = P_1 - $P_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{p1\ (1-p1)}{n1} - \frac{p2\ (1-p2)}{n2}}$
Means: 2pops Matched pairs	Known also as the "*before-and-after*" test Large sample size (CLT) if the d is not normal	$\mu_d = (\mu_1 - \mu_2)$ $\overline{d} \pm z_{\frac{\alpha}{2}} \frac{\sigma_d}{\sqrt{n}} = \overline{d} \pm z_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}}$ (if σ_d $unknown\ with\ large\ sample$) $\overline{d} \pm tz_{\frac{\alpha}{2}} \frac{\sigma_d}{\sqrt{n}} = \overline{d} \pm t_{\frac{\alpha}{2}} \frac{S_d}{\sqrt{n}}$ (if σ_d $unknown\ with\ small\ sample$
Variance σ^2	$\frac{(n-1)s^2}{X_{\alpha/2}^2} \leq \sigma^2 \leq \frac{(n-1)s^2}{X_{(1-\frac{\alpha}{2})}^2}$	$\frac{s_1^2}{s_2^2\ F(n2-1,n1-1,\frac{\alpha}{2})} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{s_1^2 F(n1-1,n2-1,\frac{\alpha}{2})}{s_2^2}$

Determination of sample size for Continuous and Discrete R.V.	with d margin of error:
	n = $\frac{z_{\alpha/2}^2\ s^2}{d^2}$
	With d margin of error for proportion:
	n = $\frac{z_{\alpha/2}^2\ \pi\ (1-\pi)}{d^2}$, $you\ may\ use\ \pi = 0.5$	

Hypothesis Testing

	1Population	2Populations
	1tailed: Ha: > or (<) Rejection region: Z > z_α or (Z < - z_α)	**2tailed: Ha: ≠** Rejection region: Z > $z_{\frac{\alpha}{2}}$ or Z < - $z_{\frac{\alpha}{2}}$
Pop mean μ	Ho: μ= μo Ha: μ≠ μo For one-sided use Z > $z_{\frac{\alpha}{2}}$ or Z < - $z_{\frac{\alpha}{2}}$) $Z = \frac{\overline{x} - \mu o}{\sigma_{\overline{x}}} \sim = \frac{\overline{x} - \mu o}{s/\sqrt{n}}$ Large sample size to invoke CLT.	Ho: μ1- μ2 = 0 Ha: μ1- μ2 ≠ 0 $Z = \frac{\overline{x1} - \overline{x2}}{\sigma_{\overline{x1}} - \sigma_{\overline{x2}}} \sim = \frac{\overline{x1} - \overline{x2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ If variances are almost the same, then use $t = \frac{\overline{x1} - \overline{x2}}{\sqrt{s_p^2\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ with table v = n1+n2-2
Population proportion (probability) ($\pi \rightarrow p = \frac{x}{n}$) Sample are sufficiently large	Ho: $\pi = \pi o$ Ha: $\pi \neq \pi o$ $Z = \frac{P - \pi o}{\sqrt{\frac{\pi o\,(1 - \pi o)}{n}}}$ Large sample size to invoke CLT.	Ho: $\pi 1 - \pi 2 = 0$ Ha: $\pi 1 - \pi 2 \neq 0$ $Z = \frac{P1 - P2}{\sigma \text{of}_{(P1 - P2)}} = \frac{P1 - P2}{\sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$ With $p = \frac{x1 + x2}{n1 + n2}$ and q = 1-p, large sample size.
Population variance	Ho: $\sigma^2 = \sigma_0{}^2$ Ha: $\sigma^2 \neq \sigma_0{}^2$ 2tailed: $X^2 > X^2_{1-\alpha/2}$ or $X^2 < X^2_{\alpha/2}$ 1tailed: $X^2 > X^2_{1-\alpha}$ (or $X^2 < X^2_{\alpha}$) With v = n-1 $X^2 = \frac{(n-1)s^2}{\sigma_0{}^2}$ Population is normal	Ho: $\frac{\sigma_1^2}{\sigma_2^2} = 1$ **1tailed:** Ha: $\frac{\sigma_1^2}{\sigma_2^2} > 1$ F= $\frac{s_1^2}{s_2^2}$, critical value F_α (n1-1, n2-1) **1tailed:** Ha: $\frac{\sigma_1^2}{\sigma_2^2} < 1$ F= $\frac{s_1^2}{s_2^2}$, critical value $1/F_\alpha$ (n2-1, n1-1) **2tailed:** Ha: $\frac{\sigma_1^2}{\sigma_2^2} \neq 1$, **F** = $\frac{Larger\ variance}{Smaller\ variance}$, critical value $F_{\alpha/2}$ (n1-1, n2-1), Populations are normal
Population mean for Pair matched	There is dependency, known also as the "*before-and-after*" test. Large sample size (CLT) if the d is not normal	Ho: μ1- μ2 = 0 Ha: μ1- μ2 ≠ 0, **Z** = $\frac{\overline{d}}{\sigma_d/\sqrt{n}} = \frac{\overline{d}}{S_d/\sqrt{n}}$

ANOVA (analysis of SS's)

As with the *t*-test, you are computing the t-statistic to test the assertion that the means of the two populations are almost the same. In a similar but extended fashion you are testing H_0: μ1= μ2 = μ3 = μ4 ...., typically with the hopes that you will be able to reject H_0 to provide evidence that the alternative hypothesis (H_a: at least on is different significantly then others) is more likely. To test H_0, you take a sample of each population; you then construct the ANOVA table. It consists of the *F*-ratio which is a ratio of two variances, called Mean Squares. In the numerator of the *F*-ratio is the $MS_{treatment}$ (i.e., $MS_{between}$) and in the denominator is the MS_{Error} (i.e., MS_{Within}). Obviously, your *F*-ratio will become larger as the $MS_{Treatment}$ becomes increasingly larger than the MS_{Error}. If F-statistic is large enough compare with critical value F_{α} (k-1, n-k), then one reject the null hypothesis.

Sources of Variation	Sum of Squares	Degrees of Freedom	Mean Squares	F-Statistic
Between	?3 = **?1-?2**	k-1	?3/k-1	?
Within	**?2**	n-k	?2/n-k	
Total	**?1** =?2 +?3	n-1		

ANOVA in Action (for demonstration purposes ONLY while saving space):

First compute **Total** Sum of Square (TSS), then compute SSW (**Within**), get SSB (**Between**) which will be readily available, after finding the first 2 SS's.

Samples from k = 3 populations (**original data**):

P1	1	2	3	mean1 = 2
P2	4	5	6	mean1 = 5
P3	7	8	9	mean1 = 8

Grand mean = (2 + 5 + 8) / 3 = **5**

Step1. TSS: Total Sum of Squares

P1	(1-**5**)^{2} =16	9	4	Sum 1 = 29
P2	(4-**5**)^{2} =1	0	1	Sum 2 = 3
P3	(7-**5**)^{2} = 4	9	16	Sum 3 = 29

TSS = 29 + 3 + 29 = **61**

d.f = (n1 + n2 + n3) -1 = (3 + 3 + 3) - 1= 9 -1 = **8**

Step2. SSW: Total Sum of Square Within

Original Data

P1	1	2	3	mean1 = 2
P2	4	5	6	mean1 = 5
P3	7	8	9	mean1 = 8

SSW

P1	(1-**2**)^{2} =1	0	1	Sum1 = 2, Variance1 = Sum1/(n1-1) = 2/(3-1) =1
P2	(4-**5**)^{2} =1	0	1	Sum2 = 2, Variance2 = Sum2/(n2-1) = 2/(3-1) = 1
P3	(7-**8**)^{2} =1	0	1	Sum3 = 2, Variance3 = Sum3/(n3-1) = 2/(3-1) = 1

SSW = 2 + 2 + 2 = **6**

d.f = (n1 -1) + (n2 -1) + (n3 -1) = 2 + 2 + 2 = **6**

Notice that we **compute variance** to check their equality condition.

Step3. SSB: Sum of Square Between

Therefore, **SSB = TSS – SSW** = 61 – 6 = **55**, **d.f.** = 8 – 6 = **2** =

Number of populations – 1 = 3 – 1 = 2

Sequential Conditions and the Check-list for Using ANOVA:

1. Samples are Random, use Runs test.

http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Randomness.htm

2. Populations are Normal (Use the Histogram), and

3. Variances are equal, use Bartlett's test:

http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm

Simple Regression Analysis

Formulas and Notations:

1. $\Sigma x / n = \bar{x}$, this is just the mean of the x values.

2. $\Sigma y / n = \bar{y}$, this is just the mean of the y values.

3. $S_{xx} = SS_{xx} = \Sigma(x(i) - \bar{x})^2 = \Sigma x^2 - (\Sigma x)^2 / n$
4. $S_{yy} = SS_{yy} = \Sigma(y(i) - \bar{y})^2 = \Sigma y^2 - (\Sigma y)^2 / n$

5. $S_{xy} = SS_{xy} = \Sigma(x(i) - \bar{x})(y(i) - \bar{y}) = \Sigma(x \cdot y) - (\Sigma x) \cdot (\Sigma y) / n$

6. Slope $b = SS_{xy} / SS_{xx}$
7. Intercept, $a = \bar{y} - b\bar{x}$

8. **Residual**(i) = **Error**(i) = y(i) – yhat(i)
9. $SSE = S_{residuals} = SS_{residuls} = SS_{errors} = \Sigma[y(i) - yhat(i)]^2 = SS_{yy} - b\, SS_{xy}$
10. Standard deviation of residuals $= s = s_e = S_{residal} = S_{errors} = [SS_{residual} / (n-2)]^{1/2}$
11. Standard error of the slope (b) $= S_b = S_{residual} / SS_{xx}^{1/2}$
12. Standard error of the intercept (a) $= S_a = S_{residual}[(SS_{xx} + n.\ ^{2}) /(n \cdot SS_{xx}]^{½}$
13. **Test of hypothesis for slope:**

H0: There is no linear relation, i.e. **slope = 0**, use a two sided t-test:

$T_{n-2}\ (\alpha/2) = b\ /\ S_b$. with n-2 d.f., at the α level.

Overall Assessment of the model:

One may use F value of ANOVA, by using the relationship between T(slope) and F tables, i.e. $\mathbf{F_{1,\ n-2}\ (\alpha) = T^2_{\ n-2}\ (\alpha/2)}$. The fit is consider "a good fit" when the F-value is at least five times of critical value of F table.

14. **The Coefficient of Determination:** The coefficient of determination is defined, and denoted by R^2:

$$R^2 = (SS_{yy} - SSE) / SS_{yy} = 1 – (SSE / SS_{yy}),\ 0 \leq R^2 \leq 1$$

The numerical value of R^2 represents the proportion of the sum of squares of deviations of the y values about their mean that can be attributed to the linear relationship between y and x. R-squares is the percentage of variance [in fact, the sum of squares] in Y accounted for by variance in X captured by the model. The reminder, 1- R^2 depends on exclusion of other factors (not X alone).

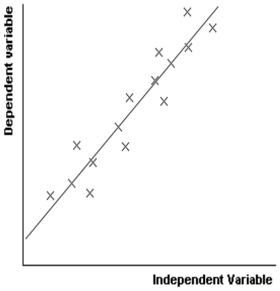
16. Prediction of y for a given x = X_0 , y-predicted = yhat = b· X_0 + a with confidence:

$$Y_p \pm S_e\ .\ t_{n-2,\ \alpha/2}\ \{\ 1 + 1/n + (X_0 – \bar{x}\ )^2 /\ S_x\}^{1/2}$$

Sequential Conditions and the Check-list for Linear Models

Almost all statistical activity of reality, including regression models have conditions (assumptions) that must be verified in order that the model has to stand the test hypotheses and for it to be able to predict accurately.

1. The dependent variable Y is a linear function of the independent variable X. This can be checked by carefully examining all the points in the **Scatter Diagram**, and ***see if it is possible to bounding them all within two parallel lines.*** Then the regression line is at the middle of these boundary lines.



Scattered Diagram to Check Linearity

2. The residuals constitute a set random variable, use the Run test:

http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/Randomness.htm

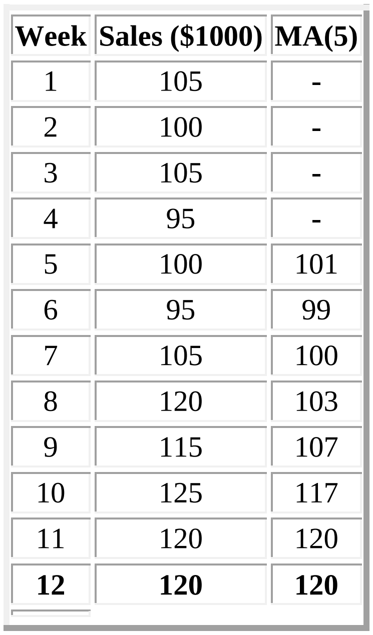
3. The distribution of the residual must be normal. Use the Histogram of error terms.

4. The residuals should have a mean equal to zero, and a constant variance. You may check this condition by dividing the residuals data into two or more groups and then computing the mean (all must be close to zero) and variance. Use the Bartlett's test for equality of variances:

http://home.ubalt.edu/ntsbarsh/Business-stat/otherapplets/BartletTest.htm

Forecasting by Moving Averages: The best-known forecasting methods is the moving averages or simply takes a certain number of past periods and add them together; then divide by the number of periods.

An illustrative numerical example: The moving average of order five are calculated in the following table.



Week	Sales ($1000)	MA(5)
1	105	-
2	100	-
3	105	-
4	95	-
5	100	101
6	95	99
7	105	100
8	120	103
9	115	107
10	125	117
11	120	120
12	**120**	**120**

Forecasting for period 13 and 14, first you **find the underlying trend** by e.g. Regression Linear fit and **then project it into future**. How good is your forecast? Exclude period 12 and forecast it, see how much error there is in there, to decide how good your real future forecast will be.

http://home.ubalt.edu/ntsbarsh/stat-data/Forecast.htm#rhowma