Chi-squared test

- Example: clinical trial of aspirin vs. placebo in the treatment of headache.

		Factor A			
		Headache	No headache	Total	Prop. w/ no headache
Factor B	Aspirin	30	70	100	0.70
	Placebo	55	55	110	0.50
Total		85	125	210	

- Could the difference in those cured by aspirin (70%) and placebo (50%) have arisen by chance?

Chi-squared test

- Expected values:

	Headache	No headache	Total
Aspirin	40.48	59.52	100
Placebo	44.52	65.48	110
Total	85	125	210

Thus, $X_c^2 = (|30-40.48|-½)^2/40.48 + (|70-59.52|-½)^2/59.52$

$+ (|55-44.52|-½)^2/44.52 + (|55-65.48|-½)^2/65.48$

$= 7.89$

Alternatively, $X_c^2 = \dfrac{(|70\text{x}55-30\text{x}55|-105)^2\,\text{x}210}{100 \text{ x } 110 \text{ x } 125 \text{ x } 85} = 7.89$

Chi-squared test

- From χ^2 table, with 1 deg. of freedom, $\chi^2_{0.005} = 7.879$. Therefore, $p<0.005$
- To approximate 95% confidence interval, $p_1=a/m$ and $p_2=c/n$.
- Standard error for difference $p_1 - p_2$ is given by: $SE(p_1-p_2) = \mathrm{sqrt}(p_1(1-p_1)/m + p_2(1-p_2)/n)$
- The 95% confidence interval for the true difference in proportions is:

$$(p_1 - p_2) \pm 1.96 \times SE(p_1 - p_2)$$

Chi-squared test

- $p_1 - p_2 = 0.20$, $SE(p_1 - p_2) = 0.066$
- 95% confidence interval: $0.20 \pm 1.96 \times 0.066$

  $0.07 - 0.33$

Contingency Tables with more than two rows or columns

- Example for the literature: Nichols et al. (1986) give the compliance with screening for colorectal cancer with respect to the method of invitation to screening.
- The three methods were: a letter with the test, a letter alone, or during a routine visit.

Contingency Tables with more than two rows or columns

Number of subjects

Method of invite	Complied	Did not comply	Total	% complied
Letter + test	3108 (3441.5)	5028 (4694.5)	8136	38.2
Letter	2468 (2648.4)	3793 (3612.6)	6261	39.4
Visit	1969 (1449.6)	1458 (1977.4)	3427	57.5
Totals	7545	10279	17824	42.3

expectation values in (   )

- Null hypothesis: compliance rate is not influenced by the method of invitation.
- $\chi^2 = \Sigma\ (O-E)^2/E$, in this case there are 6 cells, and $\chi_c^2 = 399.84$.

Contingency Tables with more than two rows or columns

- Referring to the χ^2 table with df = (r–1)(c–1) (=2), α = 0.005 → χ^2=10.597 << 399.84
- Therefore, p<0.005, there is a highly statistically significant difference in compliance rates between methods of invitation.
- Tests on r x c tables, where r or c are large, are dangerous because although the null hypothesis is clear, the alternative is not.
- If Nichols et al. had used 10 methods of invitations and only one method improved compliance, then the chi-squared test is unlikely to detect it.
- Such a test lacks statistical power.

Q1: Which one of the alternative hypotheses listed below is non-directional, and requires a two-sided statistical test?

A. H_A: "On average, a 15-year-old girl is taller than a 10-year-old girl."

B. H_A: "The drug Plavix affects average clotting time in an in vitro assay."

C. H_A: "Treatment with analgesic A will alleviate pain."

Q2: Which statistical result is most significant?

A

$$z = \frac{\bar{x} - \mu_0}{\sigma / n} = 2.5$$

B

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = 2.5$$

C

$p = 0.05$

Q3: Paired or unpaired t-test?

The migration velocity of 28 endothelial cells on rat fibronectin was measured after passage number 4.

Then the migration velocity of 28 other endothelial cells on rat fibronectin was measured after passage number 6.

H_0: The average cell migration velocity at P4 is the same as the average cell migration velocity at P6.

A. Paired t-test

B. Unpaired t-test

Is there an association between Drinking and Lung Cancer?

Suppose a case-control study is conducted to test the above hypothesis?

QUESTION: Is there a difference between the proportion of drinkers among cases and controls?

Group 1
Disease
P1= proportion of drinkers

Group 2
No Disease
P2= proportion of drinkrs

Elements of Testing hypothesis

- Null Hypothesis
- Alternative hypothesis
- Level of significance
- Test statistics
- P-value
- Conclusion

Case Control Study of Drinking and Lung Cancer

Null Hypothesis: There is no association between Drinking and Lung cancer, $P_1=P_2$ or $P_1-P_2=0$

Alternative Hypothesis: There is some kind of association between Drinking and Lung cancer, $P_1 \neq P_2$ or $P_1-P_2 \neq 0$

Based on the data in the following contingency table we estimate the proportion of drinkers among those who develop Lung Cancer and those without the disease?

		Lung Cancer		Total
		Case	Control	
Drinker	Yes	A=33	B=27	60
	No	C=1667	D= 2273	3940

ePl=33/1700 ePl=27/2300

Test Statistic

How many standard deviations has our estimate deviated from the hypothesized value if the null hypothesis was true?

$$Z = (eP1 - eP2 - 0)/[(1/n1 + 1/n2)(\sqrt{p(1-p)})]$$

$$where$$

$$p = (33+27)/(1700+2300) = 60/4000 = 3/200 = 0.015$$

$$Z = [(33/1700) - (27/2300) - 0)]/(\sqrt{(1/1700 + 1/2300)(0.015)(0.985)}$$

$$Z = 2.003$$

P-value for a two tailed test

P-value= 2 P[Z > 2.003] = 2(.024)=0.048

How does this p-value compared with α=0.05?

Since p-value=0.048 < α=0.05, reject the null hypothesis H_0 in favor of the alternative hypothesis H_A.

Conclusion:
There is an association between drinking and lung cancer.
Is this relationship causal?

Analysis of Variance (ANOVA)

- Tests for differences among three or more independent means.
- Extension of two-sample (unpaired) t-test to three or more samples.
- Test the null hypothesis that three (or more) population means are identical:

$$H_0: \mu_1 = \mu_2 = \mu_3$$

- (Alternative hypothesis is that at least one population mean differs from one of the others.)

Analysis of Variance (ANOVA)

- In general,

		Group 1	Group 2	…	Group k
Population	mean	μ_1	μ_2		μ_k
	std. dev.	σ_1	σ_2		σ_k
Sample	mean	x_1	x_2		x_k
	std. dev.	s_1	s_2		s_k
	sample size	n_1	n_2		n_k

- Assume k populations are independent and normally distributed.

Analysis of Variance (ANOVA)

- We could compare 3 population means by evaluating all possible pairs of sample means using two-sample t-test.
- For three groups, the number of required tests is:

$$\binom{3}{2} = 3$$

$$1 \leftrightarrow 2,\ 1 \leftrightarrow 3,\ 2 \leftrightarrow 3$$

Analysis of Variance (ANOVA)

- Definition: the expression "n choose x":

$$\binom{n}{x} = \frac{n!}{x!(n-x)!}$$

- Is the <u>combination</u> of n objects chosen x at a time.
- Assume variances of underlying populations are equal:

$$\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma^2$$

Analysis of Variance (ANOVA)

- Pooled estimate of common variance is:

$$s_w^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2}{n_1 + n_2 + n_3 - 3}$$

- If k=10, t-tests would become complicated, with $\binom{10}{2} = 45$ different pairwise tests.
- More importantly, many two-sample t-tests are likely to lead to an incorrect conclusion.

Analysis of Variance (ANOVA)

- Suppose that 3 population means are in fact equal and we conduct 3 pairwise tests.
- Assume the tests are independent and set the significance level for each one at 0.05.
- P(fail to reject in all 3 tests) = $(1 – 0.05)^3$

  $= 0.95^3 = 0.857$
- Therefore, the probability of rejecting H_0 in at least one test is:

P(reject in at least 1 test) = 1 – 0.857 = 0.143

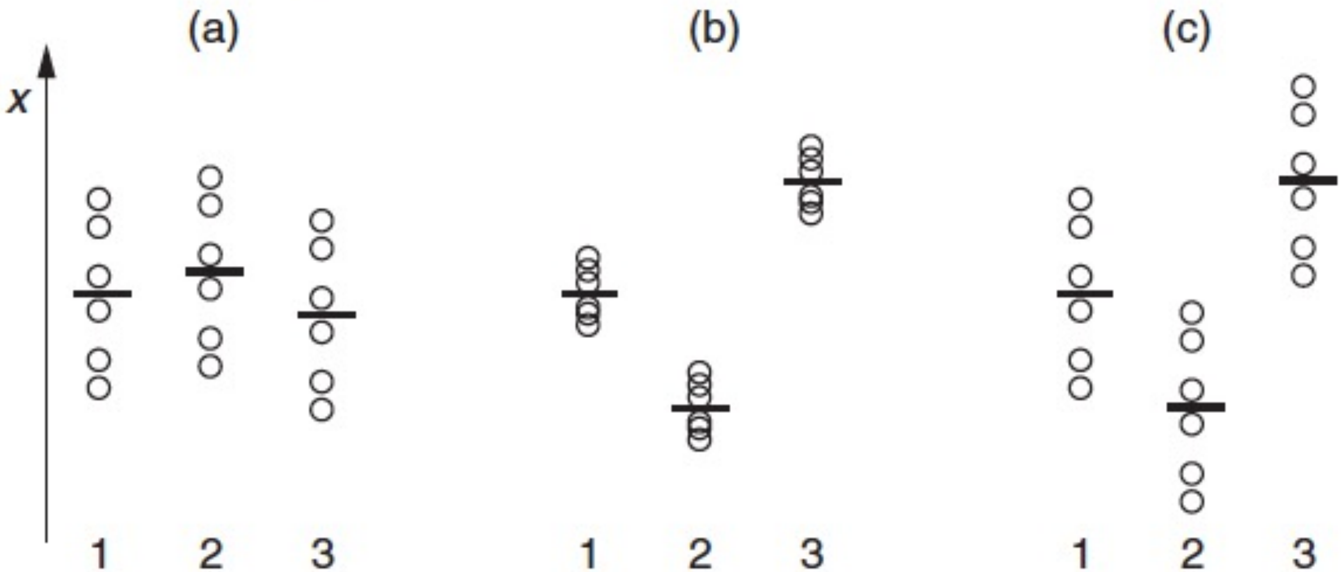
Analysis of Variance (ANOVA)

- Since the null hypothesis is true in each case, 0.143 is the overall probability of making a Type I error. Larger than 0.05!
- We need a test where the overall probability of making a Type I error is equal to some predetermined level α → one-way ANOVA.
- <u>One-way</u>: single factor/characteristic

Analysis of Variance (ANOVA)

- Two measures of variability:
  - Variations of individual values around their population means;
  - Variation of population means around the overall mean.
- If the variation within k different populations is small relative to variability among their respective means → population means are in fact different.

Three different scenarios for data variability in a multi-sample data set. The circles represent data points and the thick horizontal bars represent the individual group means.



Analysis of Variance (ANOVA)

- Null hypothesis: H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
  for a set of k populations.
- Variability of individual observations around their population means.
- Let $n = n_1 + n_2 + \ldots + n_k$
- $$s_w^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + \ldots + (n_k - 1)s_k^2}{n - k}$$
- Weighted average of k individual sample variances. **W: "within-groups" variability**

Analysis of Variance (ANOVA)

- Extent that population means vary around the overall mean.
- $$s_B^2 = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + \ldots + n_k(\overline{x}_k - \overline{x})^2}{k - 1}$$
- $(\overline{x}_i - \overline{x})^2$: squared deviation of sample means $\overline{x}_i$ from the grand mean $\overline{x}$.
- <u>Grand mean</u>: overall average of n observations that make up the k different samples.

Analysis of Variance (ANOVA)

- $\overline{x} = \dfrac{n_1\overline{x_1} + n_2\overline{x_2} + \ldots + n_k\overline{x_k}}{n}$
- B: “between-groups” variability

Analysis of Variance (ANOVA)

- Do sample means vary around the grand mean more than individual observations vary around the sample means?
- Yes? → corresponding population means

  are in fact different.
- Test statistic: $F = s_B^2/s_w^2$
- Under the null hypothesis both s_w^2 and s_B^2 estimate a common variance σ^2 → $F \sim 1$

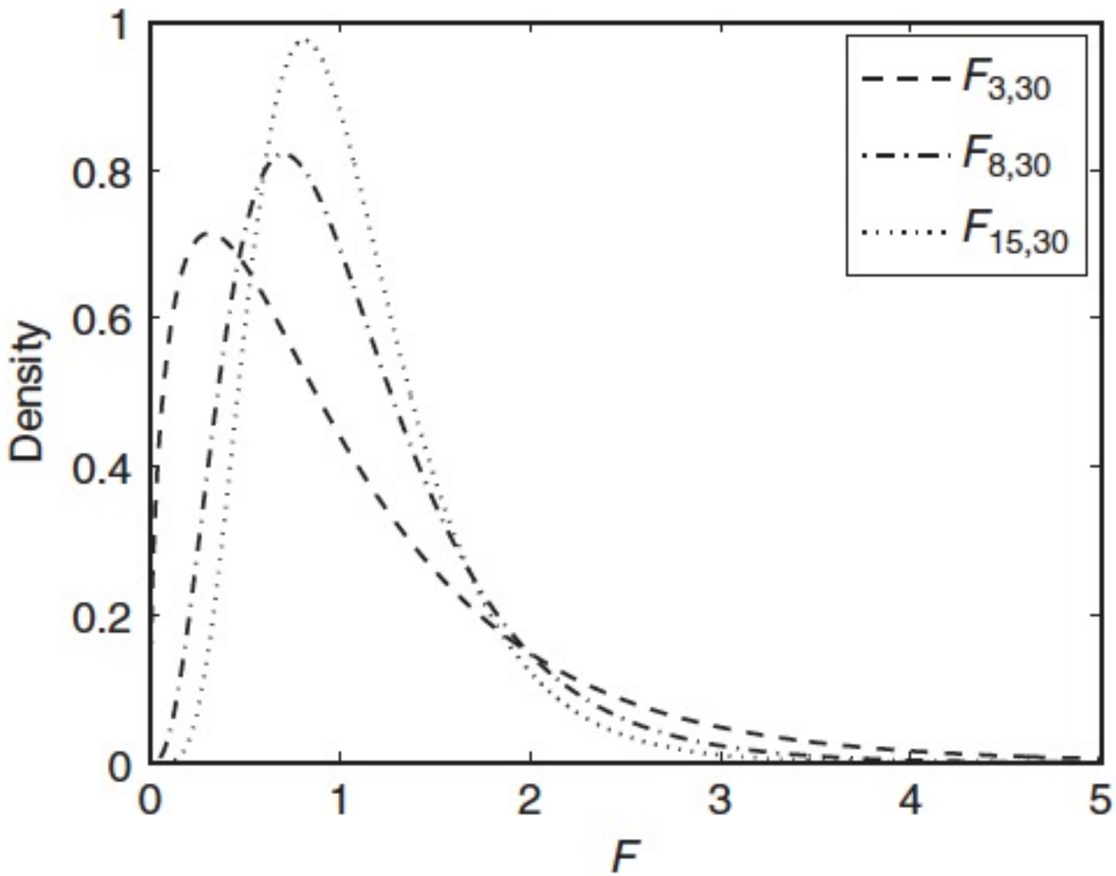
Analysis of Variance (ANOVA)

- Difference among populations → F>1
- Under H_0, the ratio F has an F-distribution with k – 1 (numerator) and n – k (denominator) degrees of freedom.
- $F_{k-1,n-k}$ or $F_{df1,df2}$
- If only 2 independent samples…

  F-test → two-sample t-test

- Different F-distribution for each possible pair of values df1 and df2.

Three *F* distributions for different numerator degrees of freedom and the same denominator degrees of freedom. The `fpdf` function in MATLAB calculates the *F* probability density.



F-distribution

- Cannot be negative.
- Skewed to the right, amount of skew depends on df's.
- Look up in table. Critical values computed for selected percentiles:
  - Upper 10.0, 5.0, 2.5, 1.0, 0.1 of distributions.
- Entry in table represents the value of Fdf1,df2 that cuts off the specified area in the upper tail of the distribution.