Chi-Square Test of Independence (based on a Contingency Table)

$$\chi^2 = \sum \frac{(Observed - E\,\mathrm{xp}\,ected)^2}{Expected}$$

$$df = (r-1)(c-1)$$

In the following contingency table estimate the proportion of drinkers among those who develop Lung Cancer and those without the disease?

		Lung Cancer		Total
		Case	Control	
Drinker	Yes	O11=33	O12=27	R1=60
	No	O21=1667	O22= 2273	R2=3940
Total		C1 = 1700	C2 = 2300	n = 4000

E11=1700(60)/4000=25.5 E12=34.5

E21=1674.5 E22=2265.5

E11=1700(60)/4000=25.5 **E12=34.5**

E21=1674.5 **E22=2265.5**

$$\chi^2_{obs} \sum_{k=1}^{k=4} \frac{(Observed - E\,\mathrm{xp}\,ected)^2}{Expected} =$$

$$\frac{\dfrac{(33-25.5)^2}{25.5} + \dfrac{(27-34.5)^2}{34.5} + \dfrac{(1667-1674.5)^2}{1674.5} + \dfrac{(2273-2265.5)^2}{2265.5}}{} = 4.0$$

Multiple Comparisons via ANOVA

- H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$
- Say we reject H_0… what now?
- Need additional tests to find where the differences lie.
- One approach: series of (k choose 2) two-sample t-tests. → increased probability of making Type I error.
- Solution: be more conservative in the individual comparisons.

Multiple Comparisons via ANOVA

- As the number of tests ↑, individual α ↓
- <u>Bonferroni correction</u>: to set overall Type I error probability at 0.05,

  α^* = 0.05/(k choose 2) for indiv. comparisons
- k=3, total of (3 choose 2)=3 tests

  → α^* = 0.05/3 = 0.0167

Multiple Comparisons via ANOVA

- H_0: $\mu_i = \mu_j$
- Calculate $t_{ij} = \dfrac{\overline{x}_i - \overline{x}_j}{\mathrm{sqrt}(s_w^2(1/n_i + 1/n_j))}$
- Note that we use all the additional info from the k samples to estimate the common variance σ^2.
- Under H_0, t_{ij} has a t-distribution with $n - k$ df.
- One drawback of Bonferroni correction: highly conservative, lacks statistical power.

Multiple Comparisons via ANOVA

- Example: Follow 3 groups of overweight males for 1 year.
- Group 1: diet, no exercise program
- Group 2: regular exercise, no diet
- Group 3: control, neither diet nor exercise.
- After 1 year, total change in body weight measured for each individual.

Multiple Comparisons via ANOVA

	Group 1	Group 2	Group 3	
n_i	42	47	42	
$\bar{x}_i$	–7.2	–4.0	0.6	in kg
s_i	3.7	3.9	3.7	in kg

- Assume independent, normally distributed data.
- Null hypothesis H_0: $\mu_1 = \mu_2 = \mu_3$
- Assume underlying population variances are equal (looks like a good assumption).

Multiple Comparisons via ANOVA

- First, estimate within-groups variance:

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

$$= 14.24 \text{ kg}^2$$

- Next, the grand mean of the data:

$$\overline{x} = \frac{n_1\overline{x}_1 + n_2\overline{x}_2 + n_3\overline{x}_3}{n_1 + n_2 + n_3}$$

$$= -3.55 \text{ kg}$$

Multiple Comparisons via ANOVA

- Then, estimate the between-groups variance:

$$s_B^2 = \frac{n_1(\overline{x}_1 - \overline{x})^2 + n_2(\overline{x}_2 - \overline{x})^2 + n_3(\overline{x}_3 - \overline{x})^2}{3 - 1}$$

$$= 646.20 \text{ kg}^2$$

- So, the test statistic is:

$$F = s_B^2 / s_w^2 = 45.38$$

- For an F-distribution with $k - 1 = 2$ and $n - k = 128$ degrees of freedom, $p < 0.001$
- Therefore, reject H_0 and conclude that the mean changes in weight for the 3 populations are not identical.

Multiple Comparisons via ANOVA

- Where are the specific differences?
- Bonferroni multiple comparisons.
- Set the overall probability for making a Type I error at 0.05: a* = 0.05/3 = 0.0167
- Diet vs. Exercise:

H_0: $\mu_1 = \mu_2$

$$t_{12} = \frac{\bar{x}_1 - \bar{x}_2}{\mathrm{sqrt}(s_w^2(1/n_1 + 1/n_2))} = -3.99$$

Multiple Comparisons via ANOVA

- For a t-distribution with $n - k = 128$ df, $p < 0.001$. → reject H_0 at the 0.0167 level of significance, conclude that μ_1 differs from μ_2.
- Diet vs. No Plan:

$$t_{13} = \frac{\bar{x}_1 - \bar{x}_3}{\mathrm{sqrt}(s_w^2(1/n_1 + 1/n_3))} = -9.47$$

- $p < 0.001$ → reject H_0, conclude that $\mu_1 \neq \mu_3$

Multiple Comparisons via ANOVA

- Exercise vs. No Plan:

$$t_{23} = \frac{\bar{x}_2 - \bar{x}_3}{\text{sqrt}(s_w^2(1/n_2 + 1/n_3))} = -5.74$$

- $p < 0.001$ → conclude that $\mu_2 \neq \mu_3$
- Therefore, all 3 population means are different from each other.

$$\mu_1 < \mu_2 < \mu_3$$

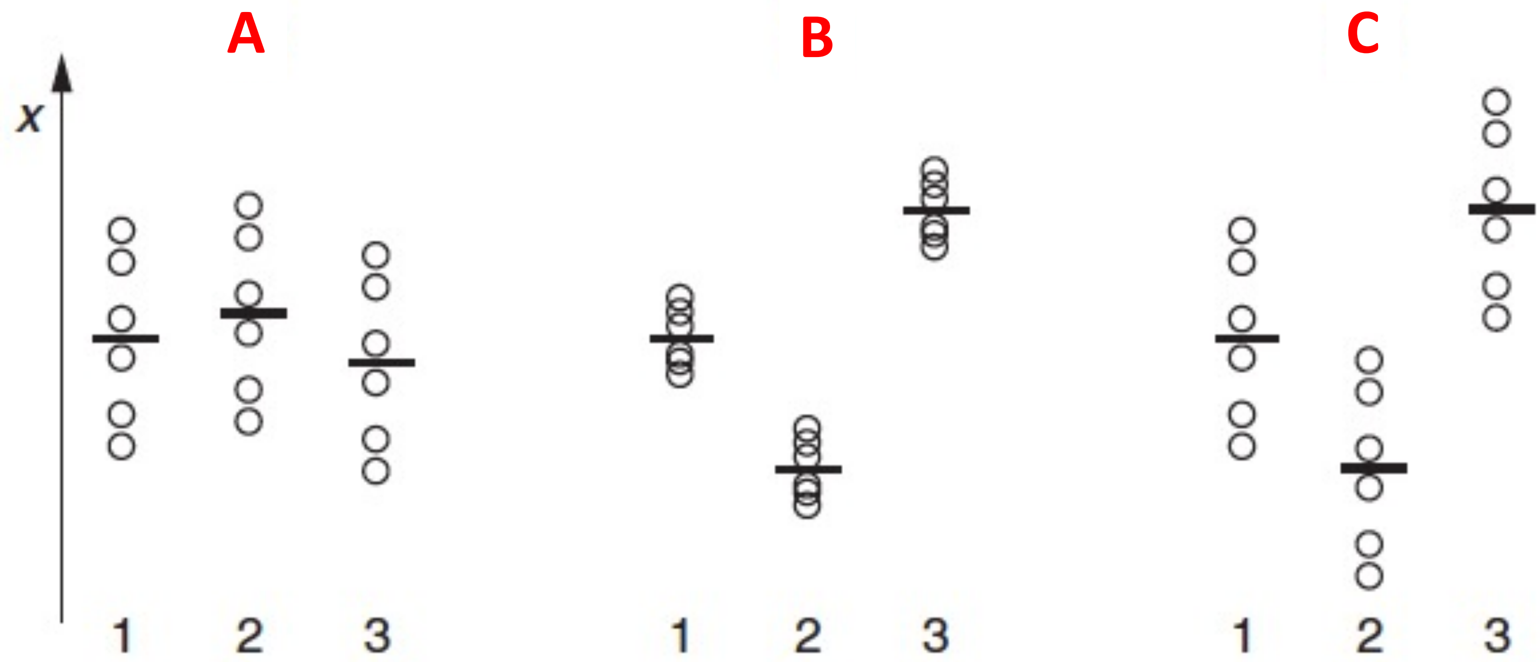
Q1: Decreasing the significance level α *decreases* the probability of making a Type II Error.

A. True
B. False

Q2: The ANOVA test produces reliable results when the following conditions hold (pick the incorrect one):

A. **The random samples are drawn from normally distributed populations.**
B. **All data points are independent of each other.**
C. **Each of the populations have the same variance σ^2.**
D. **Each of the populations have the same mean μ.**

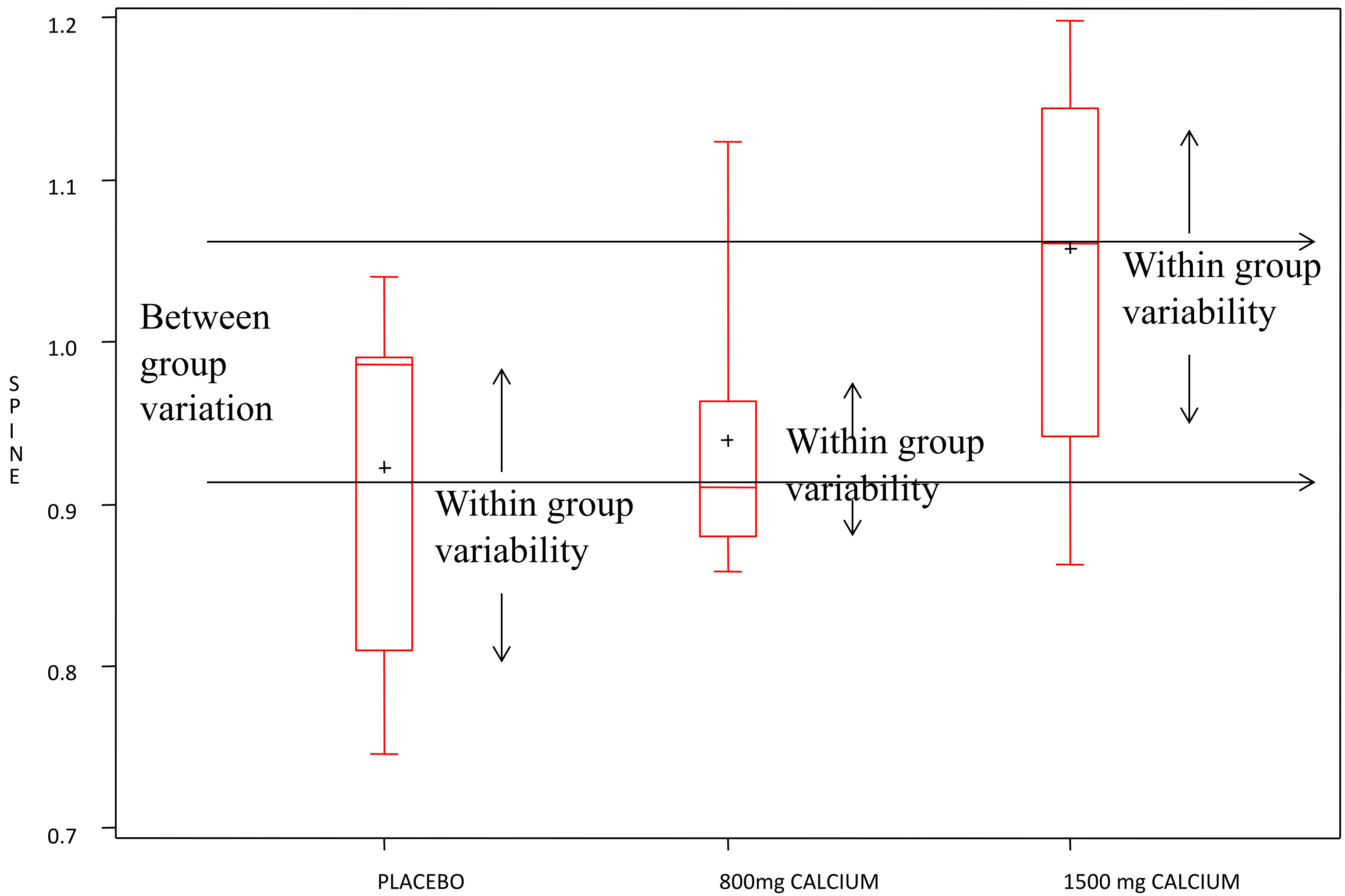
Q3: Which data set will most likely result in a rejected H_0 in an ANOVA test?



ANOVA example

- Randomize 33 subjects to three groups: 800 mg calcium supplement vs. 1500 mg calcium supplement vs. placebo.
- Compare the spine bone density of all 3 groups after 1 year.

Spine bone density vs. treatment



Group means and standard deviations

- Placebo group (n=11):
  - Mean spine BMD = .92 g/cm^2
  - standard deviation = .10 g/cm^2
- 800 mg calcium supplement group (n=11)
  - Mean spine BMD = .94 g/cm^2
  - standard deviation = .08 g/cm^2
- 1500 mg calcium supplement group (n=11)
  - Mean spine BMD =1.06 g/cm^2
  - standard deviation = .11 g/cm^2

Between-group variation.

The size of the groups.

The difference of each group's mean from the overall mean.

The F-Test

$$s^2_{between} = ns^2_{\bar{x}} = 11*(\frac{(.92-.97)^2+(.94-.97)^2+(1.06-.97)^2}{3-1}) = .063$$

$$s^2_{within} = avg\ s^2 = \frac{1}{3}(.10^2+.08^2+.11^2) = .0095$$

$$F_{2,30} = \frac{s^2_{between}}{s^2_{within}} = \frac{.063}{.0095} = 6.6$$

The average amount of variation within groups.

Each group's varian

Large F value indicates that the between group variation exceeds the within group variation (=the background noise).

Q4: Pick the most appropriate statistical test…

Are three sample means different from one another?

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

A. Two-sample *z* test

B. Paired *t* test

C. Two-sample *t* test

D. χ^2 test

E. Analysis of variance

Q5: Pick the most appropriate statistical test…

Compare the mean rolling velocity of 2500 leukocytes at a low fluid shear rate, with the mean rolling velocity of 3000 leukocytes at a high fluid shear rate.

A. Two-sample *z* test
B. Paired *t* test
C. Two-sample *t* test
D. χ^2 test
E. Analysis of variance

Q6: Pick the most appropriate statistical test…

Compare the red blood cell counts of 20 patients before and after treatment with erythropoietin.

A. Two-sample *z* test
B. Paired *t* test
C. Two-sample *t* test
D. χ^2 test
E. Analysis of variance

Q7: Pick the most appropriate statistical test…

When do fibroblasts stop spreading on fibronectin? Compare the average area of 12 cells at *t* = 30 min with the average area of 10 cells at *t* = 120 min.

A. Two-sample *z* test
B. Paired *t* test
C. Two-sample *t* test
D. χ^2 test
E. Analysis of variance

Q8: Pick the most appropriate statistical test…

Mortality rates of different groups of leukemic mice are presented in a 4 x 2 contingency table. Test for differences among the population.

A. Two-sample *z* test
B. Paired *t* test
C. Two-sample *t* test
D. χ^2 test
E. Analysis of variance

Non-normal distributions and data transformations

- Parametric tests: (e.g., t-test and ANOVA) assumes various parameters and normal distribution.
- Nonparametric tests: no assumption of population distribution is made.
- Testing if your data meets normality assumptions…

Non-normal distributions and data transformations

- **Step 1:**
- In decreasing order of robustness, either:
  1. Test that your data is not significantly deviant from a normal distribution using a test designed to do this (Kolmogorov-Smirnov, Anderson-Darling, or Shapiro-Wilks tests)
  2. Plot the data on a normal probability plot. Perfectly normal data will lie on a straight line.

Non-normal distributions and data transformations

Normal probability plot:

a. Arrange data from smallest to largest.
b. Percentile of each data value is determined.
c. From these percentiles, normal calculations are done to determine their corresponding z-scores.
d. Each z-score is plotted against its corresponding data value.

$z = (\overline{x}_1 - \overline{x}_2)/SE(\overline{x}_1 - \overline{x}_2)$ where z is distributed as a Normal distribution with μ=0 and σ=1.

Percentile $\rightarrow$ z $\rightarrow$ $\overline{x}_1$

Non-normal distributions and data transformations

- In decreasing order of robustness, either…
  1. Test the data (specialized tests)
  2. Plot on <u>normal prob plot</u>
  3. Construct a frequency histogram of your data to see whether it looks normal, i.e., symmetrical and monomodal.

Non-normal distributions and data transformations

- **Step 2:**
- If the data are <u>not</u> deviant from normality according to test in Step 1, you can now progress to using parametric statistics.
- If data are deviant from normality you must choose between the following:
  1. Use a nonparametric test. These are less sensitive, avoid if possible.
  2. Transform the data. This can help your data conform better to a normal distribution.

Non-normal distributions and data transformations

3. Use a test based on another distribution (e.g., Weibull). This is an advanced technique, used in some statistical programs.
4. Use a parametric test anyway. The test may be inaccurate so be wary of accepting results very close to $p = 0.05$. Can increase the alpha-level to 0.01 to compensate. ANOVA is a particularly robust test, resistant to distortion by non-normality.

Data Transformation

- Data transformation aims to either:
  1. Linearize (e.g., power laws, exponential functions)
  2. Normalize (make non-normal distributions appear more normal)
  3. Equalize variances (make data conform to the assumption of equal variances, "<u>homoscedatic</u>"). ANOVA assumes that sample variances are equal.

Data Transformation

Transformation	Type of data applied to
log Y or **log(Y+1)**	contagious (=aggregated) or clumped distribution, or where factors in an ANOVA are synergistic and apparently multiplicative. (use +1 if data contains zeros or numbers near 0)

Data Transformation

Transformation	Type of data applied to
arcsin($Y^{1/2}$)	Percentage (0–100%) or proportional (0 – 1) data.
sqrt(Y) or **sqrt(Y+0.375)**	Distributions where the variances are proportional to the means, i.e., they increase in unison. (Use the +0.375 form if data contain zeros or or numbers near zero)

Data Transformation

- Example of square root transformation on data where the variance increases as the mean increases.
- Data are given for the growth of tumors in three drug treatments.

Data Transformation

Control		Tumostat		Inhibin 4	
Original	**Sqr. Root**	**Original**	**Sqr. Root**	**Original**	**Sqr. Root**
17	4.12	9	3.00	5	2.24
16	4	8	2.83	4	2.00
2	1.41	3	1.73	2	1.41
1	1.00	2	1.41	1	1.00
s^2 75.34	6.92	30.25	5.02	9.00	2.75

- Original data show gross heteroscedasticity, ratio of largest variance to smallest = 8.4:1.
- Square root transformation reduces this ratio to 2.5:1.
- (for t-test, can pool s_d^2 when variances differ by ~2, more statistical power.)