Hypothesis testing: two samples

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

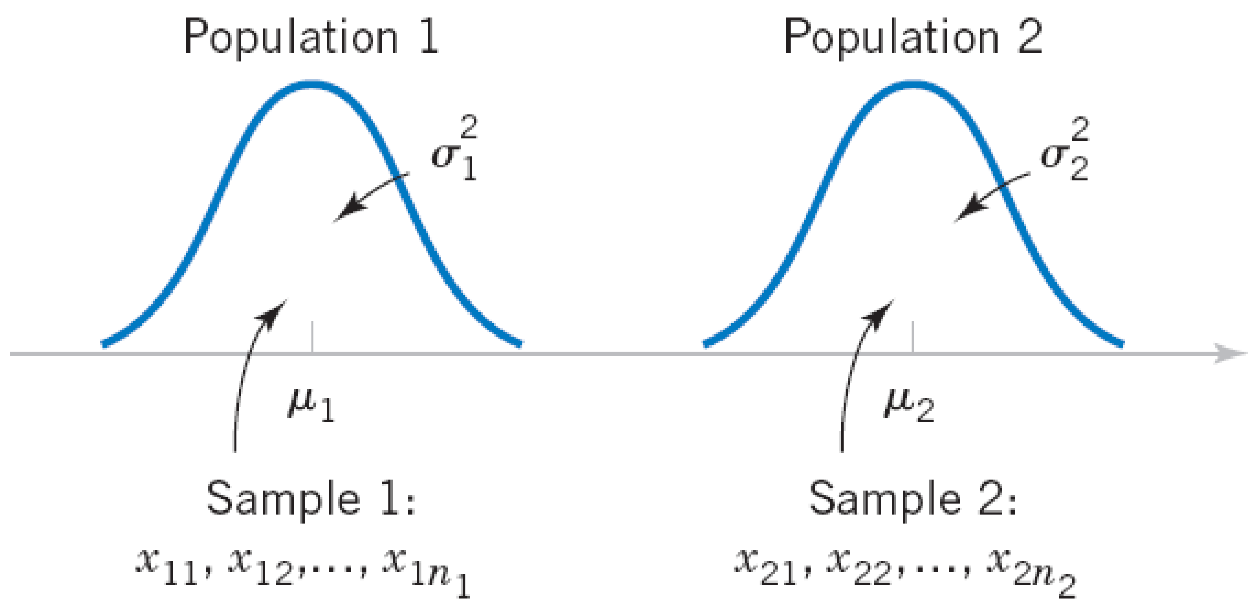


Figure 10-1 Two independent populations.

Figure 10-1 Two independent populations.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

Assumptions

1. $X_{11}, X_{12}, \ldots, X_{1n_1}$ is a random sample from population 1.
2. $X_{21}, X_{22}, \ldots, X_{2n_2}$ is a random sample from population 2.
3. The two populations represented by X_1 and X_2 are independent.
4. Both populations are normal.

$$E(\overline{X}_1 - \overline{X}_2) = E(\overline{X}_1) - E(\overline{X}_2) = \mu_1 - \mu_2$$

$$V(\overline{X}_1 - \overline{X}_2) = V(\overline{X}_1) + V(\overline{X}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

The quantity

$$Z = \frac{\overline{X}_1 - \overline{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (10\text{-}1)$$

has a $N(0, 1)$ distribution.

10-2: Inference for a Difference in Means of Two Normal Distributions, Variances Known

10-2.1 Hypothesis Tests for a Difference in Means, Variances Known

usually $\Delta_0 = 0$

Null hypothesis: $H_0\colon \mu_1 - \mu_2 = \Delta_0$

Test statistic:
$$Z_0 = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}} \qquad (10\text{-}2)$$

Alternative Hypotheses	P-Value	Rejection Criterion For for Fixed-Level Tests
$H_1\colon \mu_1 - \mu_2 \neq \Delta_0$	Probability above $\lvert z_0 \rvert$ and probability below $-\lvert z_0 \rvert$, $P = 2[1 - \Phi(\lvert z_0 \rvert)]$	$z_0 > z_{\alpha/2}$ or $z_0 < -z_{\alpha/2}$
$H_1\colon \mu_1 - \mu_2 > \Delta_0$	Probability above z_0, $P = 1 - \Phi(z_0)$	$z_0 > z_\alpha$
$H_1\colon \mu_1 - \mu_2 < \Delta_0$	Probability below z_0, $P = \Phi(z_0)$	$z_0 < -z_\alpha$

10-2.1 Hypotheses Tests on the Difference in Means, Variances Unknown

Case 2: $\sigma_1^2 \neq \sigma_2^2$

If H_0: $\mu_1 - \mu_2 = \Delta_0$ is true, the statistic

$$T_0^* = \frac{\overline{X}_1 - \overline{X}_2 - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}} \qquad (10\text{-}15)$$

is distributed as t-distribution with degrees of freedom given by

$$v = n_1 + n_2 - 2,$$

or more generally

Multiple null hypotheses: Bonferroni correction

- What if you have **m** independent null hypotheses? Say you have **m=25,000 genes** in a genome?
- What is the probability that at least one of the null-hypotheses will be shown to be false at significance threshold α_1?
- Answer:
  Family-Wise Error Rate
  or $FWER=1-(1-\alpha_1)^m \approx m\alpha_1$
- If $m=20$ and $\alpha_1=0.05$,
  $FWER= 0.6415$
- If you want to get $FWER< \alpha$, use
  $\alpha_1= \alpha/m$

Carlo Emilio Bonferroni
(1892 –1960)
Italian mathematician
who worked on
probability theory.

Did you know that M&M's® Milk Chocolate Candies are supposed to come in the following percentages: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown?

http://www.scientificameriken.com/candy5.asp

"To our surprise M&Ms met our demand to review their procedures in determining candy ratios. It is, however, noted that the figures presented in their email differ from the information provided from their website (http://us.mms.com/us/about/products/milkchocolate/). An email was sent back informing them of this fact. To which M&Ms corrected themselves with one last email:

In response to your email regarding M&M'S CHOCOLATE CANDIES

Thank you for your email.
On average, our new mix of colors for M&M'S® Chocolate Candies is:

M&M'S® Milk Chocolate: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown.

M&M'S® Peanut: 23% blue, 23% orange, 15% green, 15% yellow, 12% red, 12% brown.

M&M'S® Kids MINIS®: 25% blue, 25% orange, 12% green, 13% yellow, 12% red, 13% brown.

M&M'S® Crispy: 17% blue, 16% orange, 16% green, 17% yellow, 17% red, 17% brown.

M&M'S® Peanut Butter and Almond: 20% blue, 20% orange, 20% green, 20% yellow, 10% red, 10% brown.

Have a great day!

Your Friends at Masterfoods USA
A Division of Mars, Incorporated

How to accept or reject the null hypothesis that these probabilities are correct from a finite sample?

Pearson chi^2 Goodness of Fit Test

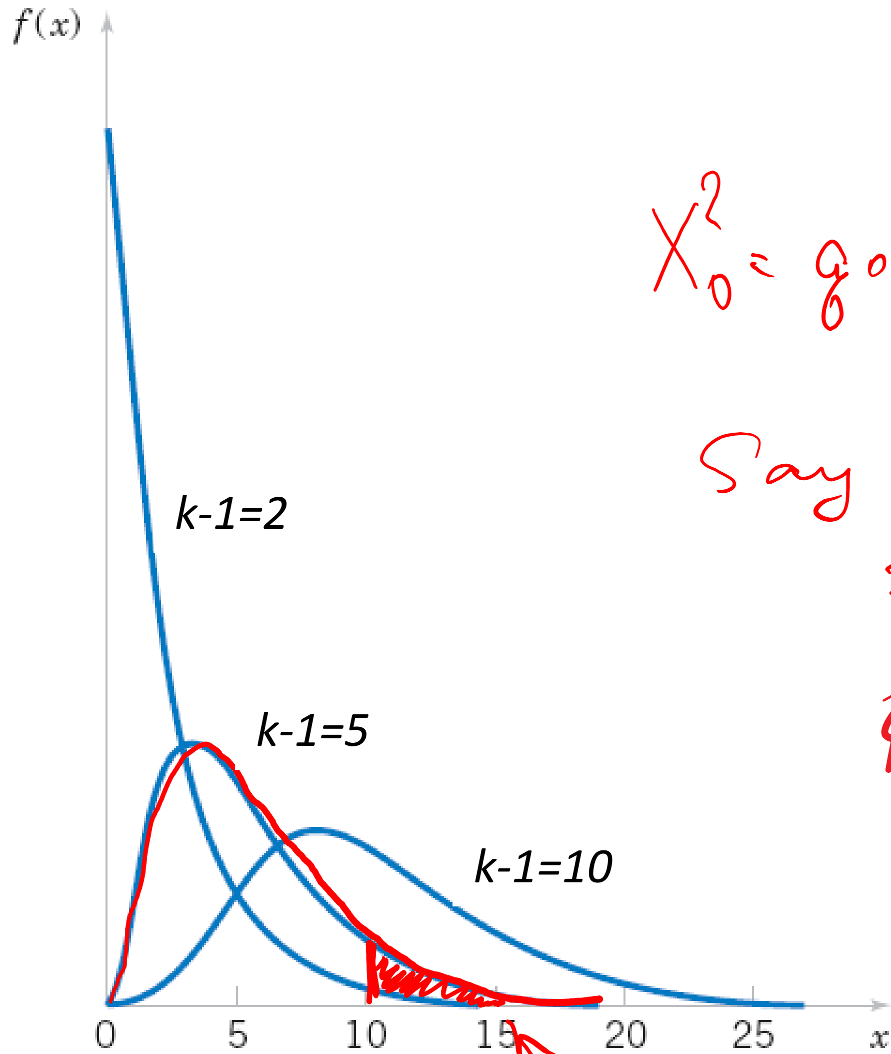
- Assume there is a sample of size n from a population with k classes (e.g. 6 M&M colors)
- Null hypothesis H_0: class i has frequency f_i in the population
- Alternative hypothesis H_1: some population frequencies are inconsistent with f_i
- Let O_i be the observed number of sample elements in the ith class and $E_i = n\, f_i$ be the expected number of sample elements in the ith class.
- Group any bin with E_i <3 with
  a) if numerical value of i is important, group it with its neighbor (k=i-1 or k=i+1) which has the smallest E_k until E_{group} >=3;
  b) If numerical value of i is irrelevant, group together all E_i <3 bins until E_{group} >=3
- The test statistic is

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \qquad (9\text{-}47)$$

P-value is calculated based on the chi-square distribution with k-1 degrees of freedom:

P-value = Prob(H_0 is correct) =1-CDF_chi-squared(X_0^2, k-1)

chi^2 Goodness of Fit Test is a ***one-sided*** hypothesis



$$X_0^2 = gof = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Say $X_0^2 = 10$

For M&M

$k = 6 \rightarrow k-1 = 5$

X_0^2

p-value that null hypothesis is correct

M&M group exercise

- DO NOT EAT CANDY BEFORE COUNTING IS FINISHED! THEN, **PLEASE, DO**.
- We will be testing three null hypotheses one after another:
  - M&M official data: 24% blue, 20% orange, 16% green, 14% yellow, 13% red, 13% brown
  - Website (fan collected) data from http://joshmadison.com/2007/12/02/mms-color-distribution-analysis: 18.36% blue, 20.76% orange, 18.44% green, 14.08% yellow, 14.20% red, 14.16% brown
  - Uniform distribution: 1/6~16.67% of each candy color
- You will estimate P-values for each one of these null hypotheses
- Hints: O_i – is the observed # of candies of color i; calculate the expected # E_i=(# candies in your sample)*f_i

Use **1-chi2cdf(X0squared, 5)** for P-value

$$X_0^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

M&M matlab exercise

- observed=mm_table(group,:); group % use when analyzing one group
- f_mm=[0.24,0.2,0.16, 0.14, 0.13,0.13];
- f_u=1./6.*ones(1,6);
- f_website=[18,21,18,14,14,14,14];
- f_website=f_website./sum(f_website);
- %p_website=[0.1836, 0.2076, 0.1844, 0.1408, 0.1420, 0.1416]
- %p_u=[0.1500, 0.2200, 0.2100, 0.1200, 0.1600, 0.1500];
- n=sum(observed)
- expected_u=n.*f_u;
- expected_mm=n.*f_mm;
- expected_website=n.*f_website;
- gf_mm=0; gf_u=0; gf_website=0;
- for m=1:6;
-     gf_mm=gf_mm+(observed(m)...
-         -expected_mm(m)).^2./expected_mm(m);
-     gf_u=gf_u+(observed(m)-expected_u(m)).^2./expected_u(m);
-     gf_website=gf_website+(observed(m)...
-         -expected_website(m)).^2./expected_website(m);
- end;
- disp('goodness of fit of MM ='); disp(num2str(gf_mm));
- disp('p-value of MM ='); disp(num2str(1-chi2cdf(gf_mm,5))); disp(' ');
- disp('goodness of fit of website ='); disp(num2str(gf_website));
- disp('p-value of MM ='); disp(num2str(1-chi2cdf(gf_website,5))); disp(' ');
- disp('goodness of fit of uniform ='); disp(num2str(gf_u));
- disp('p-value of uniform='); disp(num2str(1-chi2cdf(gf_u,5)));

Statistical tests of independence

How to **test the hypothesis** if multiple sample are drawn from the same population?

- Table: samples (Student groups) – rows, classes (M&M colors) – columns
- Test if color fractions are **independent** from group
- ***P(Group 1 and Color = green) = P(Group 1)*P(Color green)***
- Compute for all groups/colors 6*4=24 in our case

  $E_{green}(group\ 1) = n_{tot} * (group1/n_{tot}) * (green/n_{tot})$

- $\chi^2 = \sum_{groups\ \&\ colors}^{n_{tot}} \frac{(O_{color}(group) - E_{color}(group))^2}{E_{color}(group)}$
- # degrees of freedom=(colors-1)*(groups-1)

- M&M exercise Spring 2024
- Was the M&M box from Costco well mixed?
  Let's compare the first two groups' data

Title	Blue	Orang	Green	Yellov	Red	Brow	Sampl	Origin
group 1	29	22	34	45	41	14	185	Costco
group 2	30	28	25	43	44	27	197	Costco
all Costco	59	50	59	88	85	41	382	
							0	
group 3	44	30	52	10	50	27	213	Schnucl
group 4	53	31	58	17	41	30	230	Schnucl
all Schnuc	97	61	110	27	91	57	443	

- Using $\chi^2 = \sum_{groups\ \&\ colors}^{24} \frac{(O_{color}(group) - E_{color}(group))^2}{E_{color}(group)}$

with # degrees of freedom (colors-1)*(groups-1)

Find P-value of null hypothesis H_0 that
samples are independent from each other

Was the Costco box well mixed?

- `clear mm_table`
- `mm_table=mm_table_all(1:2,:);`
- `ngroups=2;`
- `ncolors=6;`
- `sumt=sum(sum(mm_table))`
- `sum_color=sum(mm_table, 1)`
- `sum_group=sum(mm_table, 2)`
- `mm_exp=kron(sum_group,sum_color)./sumt`
- `gof=sum(sum((mm_table-mm_exp).^2./mm_exp))`
- `P_value_gof=1-chi2cdf(gof, (ngroups-1)*(ncolors-1))`
- `%gof = 6.0121; P_value_gof = 0.3050`
- The null model that samples are independent is **not rejected → The Costco box was well mixed!**

Batch effect

Does color composition vary between Costco and Schnucks

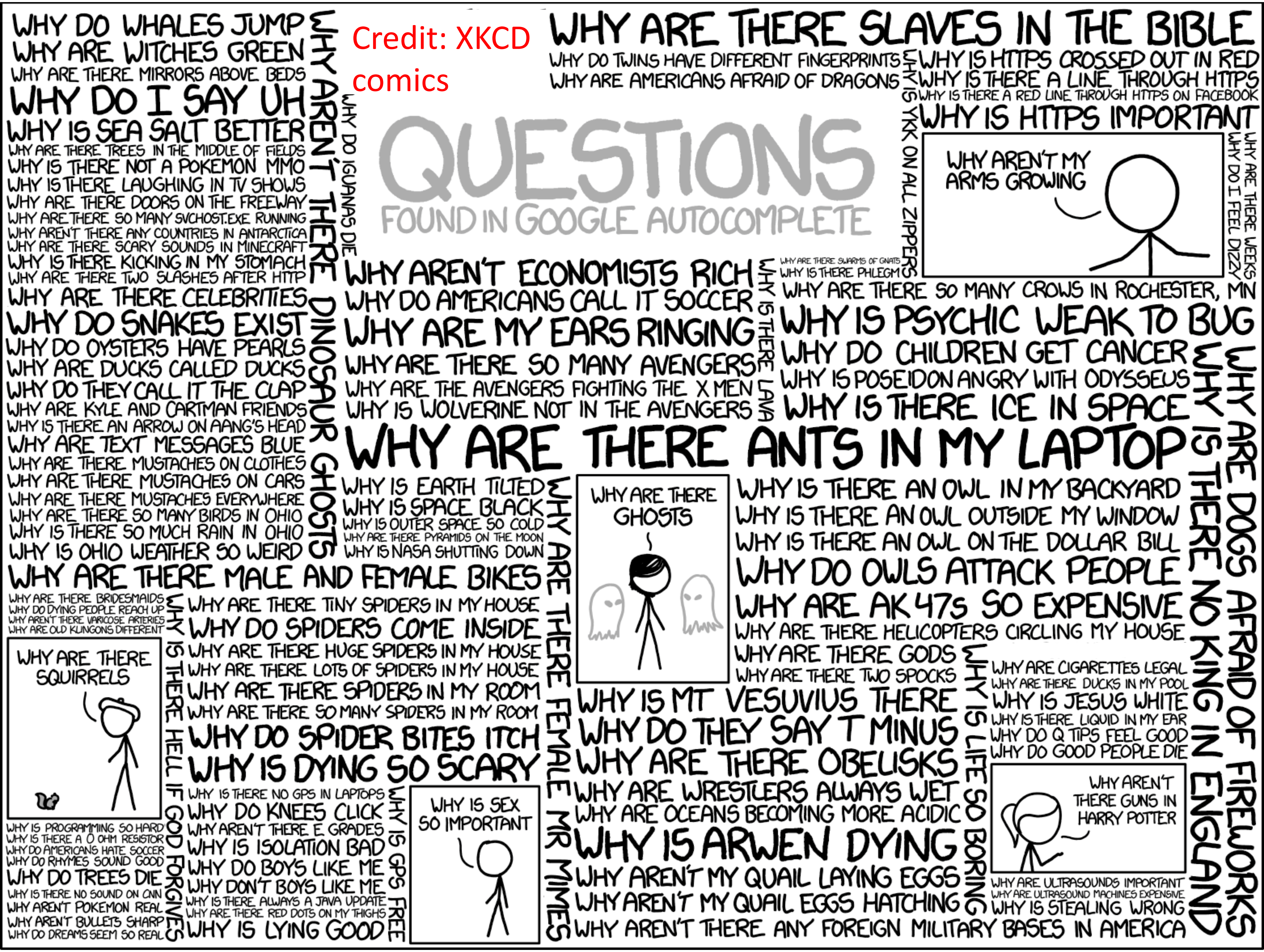
- Costco: 59 50 59 88 85 41
- Schnucks: 97 61 110 27 91 57
- Test if they are significantly different from each other:
- Same statistical independence test: ngroups=2; ncolors=6;
- Results:

  Goodness of Fit = 56.7101

  P-value = 5.8028e-11
- Batch effect is highly statistically significant! Costco and Schnucks do nor represent the same population

Do Costco (groups 1 and 2) and Schnucks (groups 3 and 4) data come from the same population (factory?)

- `clear mm_table`
- `mm_table(1,:)=sum(mm_table_all(1:2,:));`
- `mm_table(2,:)=sum(mm_table_all(3:4,:));`
- `ngroups=2;`
- `ncolors=6;`
- `sumt=sum(sum(mm_table))`
- `sum_color=sum(mm_table, 1)`
- `sum_group=sum(mm_table, 2)`
- `mm_exp=kron(sum_group,sum_color)./sumt`
- `gof=sum(sum((mm_table-mm_exp).^2./mm_exp))`
- `P_value_gof=1-chi2cdf(gof, (ngroups-1)*(ncolors-1))`
- `% Goodness of Fit = 56.7101`
- `% P-value = 5.8028e-11`
- The null model that samples are independent is **rejected**
- **Costco and Schnucks get candy from different factories**



Credit: XKCD comics