Consider the dataset x = <4,2,6,1,6,5,1,1,3,1,3,2,4,6,6> n = 15. We need a "test statistic" that measured the departure from the null hypothesis. Let's begin by tabulating the frequencies for each face of the diet:

Face	1	2	3	4	5	6	
Counts	4	1	3	2	1	4	15
Expected counts given H_0	2.5	2.5	2.5	2.5	2.5	2.5 $= \frac{1}{6} \cdot 15$	15

Denote the rv of the counts $O_1, O_2, \ldots, O_6$ where "O" means "observed"

Call the expected counts $E_1, E_2, \ldots, E_6$

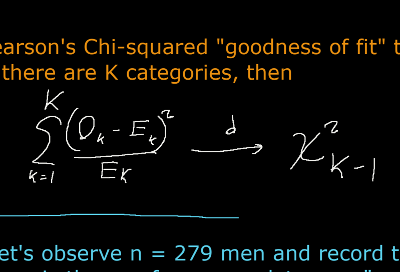
We need one value (a metric) that gauges departure from the null.

$$\hat{\phi} = |O_1 - E_1| + |O_2 - E_2| + \ldots + |O_6 - E_6|$$

This is a reasonable metric. But now we need its sampling distribution. Which we don't know... Consider...

$$\hat{\phi} = \frac{(O_1 - E_1)^2}{E_1} + \frac{(O_2 - E_2)^2}{E_2} + \ldots + \frac{(O_6 - E_6)^2}{E_6} \xrightarrow{d} \chi^2_5$$

$\Rightarrow \hat{\phi} \sim \chi^2_5$ and then you can run the test



$$\hat{\phi} = \frac{(4-2.5)^2}{2.5} + \frac{(1-2.5)^2}{2} + \ldots + \frac{(4-2.5)^2}{2.5} = 3.8 \in RET$$

$\Rightarrow$ Retain H_0

There is not sufficient evidence that our die is unfair

Pearson's Chi-squared "goodness of fit" test (Karl Pearson, 1900). If there are K categories, then

$$\sum_{k=1}^{K} \frac{(O_k - E_k)^2}{E_k} \xrightarrow{d} \chi^2_{K-1}$$

Proof is beyond scope of course. Need more advanced prob. theory than in MATH 368.

Let's observe n = 279 men and record their hair and eye color. Here is the raw frequency data as a "contingency table" or "cross tabulation"

Eye Color

Hair Color	Brown	Blue	Hazel	Green	
Black	$O_{11} = 32$	$O_{12} = 11$	10	3	$n_{HB} = 56 = n_{1\cdot}$
Brown	$O_{21} = 53$	50	25	15	$n_{HO} = 143 = n_{2\cdot}$
Red	10	10	7	7	$n_{HR} = 34 = n_{3\cdot}$
Blond	3	30	5	8	$n_{HL} = 46 = n_{4\cdot}$
	$n_{EB} = 98 = n_{\cdot 1}$	$n_{EL} = 101 = n_{\cdot 2}$	$n_{EH} = 47 = n_{\cdot 3}$	$n_{EG} = 33 = n_{\cdot 4}$	$n = 279$

$r = 4$, $c = 4$

What are we going to test? We would like to show an "association" or "relationship" between hair and ey color (in men) AKA hair and eye color are dependent (alternative hypothesis). Thus, the null hypothesis is that hair and eye color are independent. What are the parameters? Let θ represent true probabilities.

H_0: $\theta_{EB\&HB} = \theta_{EB}\theta_{HB}$ and $\theta_{EL\&HB} = \theta_{EL}\theta_{HB}$ and

$\ldots \ldots \quad \theta_{EG\&HL} = \theta_{EG}\theta_{HL}$

the number of equations is the number of cells in the cross tab = rc.

H_a: at least one of the rc above is unequal

i.e. the probability of landing in a cell = probability of landingin that cell's row times the probability of landing in that cell's column

How to run this test? We need a metric that captures departure from the null hypothesis. We first compute all the expected counts under the null. So an entire crosstab of expected counts.

eye color

hair color	EB	EL	EZ	EG	
HB	$E_{11} = n\hat{\theta}_{HB}\hat{\theta}_{EB} = 19.66$				$\hat{\theta}_{HB} = \frac{n_{HB}}{n} = \frac{56}{279} = .201$
HO					
HR					
HL					
	$\hat{\theta}_{EB} = \frac{n_{EB}}{n} = \frac{98}{279} = .351$				

$$\hat{\phi} = \underbrace{\frac{(O_{11} - E_{11})^2}{E_{11}} + \ldots + \frac{(O_{44} - E_{44})^2}{E_{44}}}_{rc = 16 \text{ terms}} = \sum_{i=1}^{4}\sum_{j=1}^{4} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 41.28$$

What is the sampling distribution of this statistic?

$$\hat{\phi} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(O_{ij} - n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j})^2}{n\hat{\theta}_{i\cdot}\hat{\theta}_{\cdot j}}$$

$$= \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{\left(N_{ij} - n\frac{N_{i\cdot}}{n}\frac{N_{\cdot j}}{n}\right)^2}{n\frac{N_{i\cdot}}{n}\frac{N_{\cdot j}}{n}} \xrightarrow{d} \chi^2_{(r-1)(c-1)}$$

This is the "chi-squared test of independence" which is a goodness of fit test where the fit you want is the one that is given by independence of the two categorial random variables. There is a third test, the "chi-squared test of homogeneity". We won't cover it.

Let's now run our test at alpha = 5%. (r-1)(c-1) = 9. $F^{-1}_{\chi^2_9}(.95) = 16.91$

$\Rightarrow \hat{\phi} = 41.28 > 16.91 \Rightarrow$ Reject H_0. Hair color and eye color are dependent.

Midterm II ↑

Final Exam ↓

This class has focused mainly on the three goals of statistical inference: estimation, testing and confidence sets. We will continue with this but now do some "meta" concepts.

Usually you're given a dataset x_1, x_2, ..., x_n and you assume a DGP, define one inferential targets θ, estimate its value using a point estimator $\hat{\theta}$, run a test by seeing if that estimate is sufficiently far from the θ_0 in the null hypothesis, or create a confidence set of possible θ's.

How do you assume a DGP? Sometimes you really know e.g. a coin flip must be Bernoulli, a die is uniform discrete, etc. But you usually you don't know! What is the DGP for wind speeds at JFK airport? What is the DGP for survival times for rats? What is the DGP for the percentage movements of the stock market? Maybe they're very complicated! (You can mostly always do statistical inference for the mean via the CLT).

What if we want to look at data and get its DGP? That's probably too hard. What if we have a set of candidate DGP models and see which one fits the best? That's doable. This is called "positing DGP's for data" or simply "model fitting".

Assum M DGP models (models) m = 1, 2, ..., M and provide a protocol to (a) select the best fitting model m_* and (b) provide scores to each of the candidate models. Goal (a) is also called the "model selection" problem. It is a core, fundamental, foundational problem in all of the scientific endeavor.

Take an example of gravity. When classical physicists were positing formulas / models for gravity, they came up with:

$m = 1 \quad F = G\frac{m_1 m_2}{r^2}$ Newton's law

$m = 2 \quad F = G_1\frac{m_1 m_2}{r^2} + G_2\frac{m_1 m_2}{r^3}$ Newton's extension to his law

$m = 3 \quad F = G_1\frac{m_1 m_2}{r^2}e^{-G_2 r}$ Laplace's extension

⋮ more...

which is the best? But they're all wrong since Einstein's equation fits better using his theory of general relativity. Likely Einstein will be proven to be wrong too since there are huge open problems in physics today. In MATH 342W, you'll pick a model purely based on how it fits the data (atheoretical). Our protocol here will be more theoretical.

$$DGP_1 : \overset{iid}{\sim} f_1(x_i; \theta_{11}, \ldots \theta_{1K_1}) = \mathcal{L}_1(\theta_{11}, \ldots \theta_{1K_1}; x_i)$$

$$DGP_2 : \overset{iid}{\sim} f_2(x_i; \theta_{21}, \ldots, \theta_{2K_2}) = \mathcal{L}_2(\theta_{21}, \ldots, \theta_{2K_2}; x_i)$$

$$\vdots$$

$$DGP_M : \overset{iid}{\sim} f_M(x_i; \theta_{M1}, \ldots, \theta_{MK_M}) = \mathcal{L}_M(\theta_{M1}, \ldots, \theta_{MK_M}; x_i)$$

Note: the number of parameters K_m may be different for each of the M candidate models.

How do we do goal (a) model selection - pick best model? How about do the naive thing: pick the model with the highest likelihood? That is "who fits the data the best"?

$$m_* := \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \left\{ \prod_{i=1}^{n} \mathcal{L}_m(\theta_{m1}, \ldots \theta_{mK_m}; x_i) \right\}$$

$$= \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \ell_m(\theta_{m1}, \ldots, \theta_{mK_m}; x_i) \right\}$$

Are we done? Can we do this? We can't do this since we don't know any of the values of the θ's for any model.

Why not replace the θ's with their estimates? And since the MLE's are really good estimates, let's do that:

$$m_* \approx \underset{m \in \{1,\ldots,M\}}{\operatorname{argmax}} \left\{ \sum_{i=1}^{n} \ell_m\left(\hat{\theta}^{MLE}_{m1}, \ldots, \hat{\theta}^{MLE}_{mK_m}; x_i\right) \right\}$$