Tyler Maule

$$\left(\begin{bmatrix}X_{11}\\ \vdots \\ X_{1d}\end{bmatrix}, \dots, \begin{bmatrix}X_{K1}\\ \vdots\\ X_{Kd}\end{bmatrix}\right) \sim \text{multinomial}\left(n, \begin{bmatrix}\pi_{11}\\ \vdots\\ \pi_{1d}\end{bmatrix}, \dots, \begin{bmatrix}\pi_{K1}\\ \vdots\\ \pi_{Kd}\end{bmatrix}\right)$$

X_{11}	X_{21}	X_{31}	$\cdots$	X_{K1}	X_{+1}
X_{12}	X_{22}	$\cdots$		X_{K2}	X_{+2}
$\vdots$	$\vdots$	$\ddots$		$\vdots$	
X_{1d}	X_{2d}	$\cdots$		X_{Kd}	X_{+d}
X_{1+}	X_{2+}			X_{K+}	X_{++}

π_{11}		π_{K1}	π_{+1}
		$\vdots$	
π_{1d}		π_{Kd}	π_{+d}
π_{1+}		π_{K+}	π_{++}

we know that $\sum\sum X_{ij} = n = X_{++}$ $\quad \sum\sum \pi_{ij} = \pi_{++} = 1$

we know that marginal dist. of $y_{ij} \sim \text{binomial}(n, \pi_{ij})$

we know the sum of two binomial variables $A \sim \text{bin}(n, p_a)$ and $B \sim \text{bin}(n, p_b)$ that is $A + B = C \sim \text{bin}(n, p_a + p_b)$ (w/ same n & disjoint "successes" on same sample space)*

so $X_{j+} = \sum_{i=1}^{d} X_{ji} \sim \text{binomial}(n, \sum_{i=1}^{d} \pi_{ji}) = \text{binomial}(n, \pi_{j+})$

then since we know that $X_{j+} \sim \text{binomial}(n, \pi_{j+})$

And by the definition (Agresti pg. 6): if $y_{ij} = 1$ if trial i has outcome in category j and $y_{ij} = 0$ otherwise $\Rightarrow$ $\vec{y_i} = [y_{i1}, y_{i2}, y_{i3}, \dots, y_{ic}]$ represents a multinomial trial ... etc.*

$(X_{1+}, X_{2+}, \dots, X_{K+}) \sim \text{multinomial}(n, \pi_{1+}, \pi_{2+}, \dots \pi_{K+})$

* further explanation on following page

Tyler Maule — AQA HOMEWORK I

pt. 1 continued, & preface for question two

Additional/Alternate method for showing sum of disjoint binomial or bernoulli trials added up take the form of a binomial...

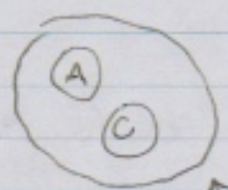
$$\text{MGF of } \mu_+(t_1, \dots, t_k) = E\left[\exp\left(\sum_{u=1}^{K} t_i X_{u+}\right)\right] = E\left[\exp\left(\sum_{u=1}^{K} t_u \sum_{i=1}^{d_u} X_{ui}\right)\right]$$

$$= \left[\sum_{u=1}^{K}\sum_{i=1}^{d_u} \pi_{ui} \exp(t_i)\right]^m$$

$$= \left[\sum_{u=1}^{K} \pi_{u+} \exp(t_u)\right]^m$$

which gives the MGF of a multinomial.

So a vector of sums of binomial or bernoulli trials w/ disjoint "successes" on the same sample space is indeed a multinomial trial. (in the same fashion, the sum of binomial RVs is still binomial if they (1) have same # of trials and (2) have disjoint "successes" on the same sample space)



n trials on A

i.e. $X \sim bin(n, \pi_a)$ $\quad Y \sim bin(n, \pi_c)$

$(X+Y) \sim bin(n, \pi_a + \pi_c)$

1.2

Let $(X_1, \dots, X_6) \sim \text{multinomial}(n, \pi_1, \dots, \pi_6)$

Show that $(X_1 + X_3, X_2, X_4 + X_5) \sim \text{multinomial}(n, \pi_1, \dots, \pi_5; \sum_{i=1} \pi_i \leq 1)$

by def. multinomial dist, $\sum_{i=1}^{c} \pi_i = \pi_{++} = 1$, where in this case $c = 6$

by def. the marginal distribution of a multinomial dist is binomial, $X_i \sim (n, \pi_i)$

so by the prop. binomial dist., $A \sim bin(n, \pi_a)$, $B \sim bin(n, \pi_b)$
$\Rightarrow A + B = C \sim bin(n, \pi_a + \pi_b)$

$X_1 + X_3 \sim bin(n, \pi_1 + \pi_3)$
$X_2 \sim bin(n, \pi)$
$X_4 + X_5 \sim bin(n, \pi_4 + \pi_5)$
$X_6 \sim bin(n, \pi_6)$

Since $(x_1 + x_3)$ by def. binomial denotes # of trials corresponding to outcomes $j = 1$ and $j = 3$

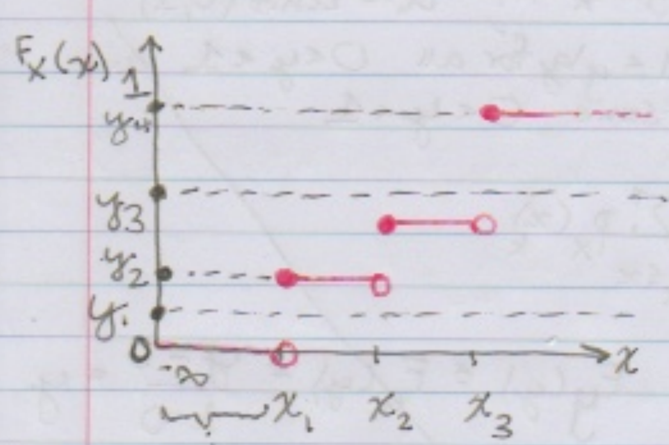
so too with x_2 for $j = 2$, $x_4 + x_5$ for $j = 4$ and $j = 5$

but as $\pi_{++} = \sum_{i=1}^{6} \pi_i = 1$, necessarily $\sum_{i=1}^{5} \pi_i \leq 1$

So by definition $(X_1 + X_3, X_2, X_4 + X_5) \sim \text{multinomial}(\cdots\cdots; \sum_{i=1}^{5} \pi_i \leq 1)$ as desired

$y = F_X(x)$ (1.3)

show $F_y(y) \le y \ \forall\, 0<y<1$ w/ some equality



In other terms:

$$P[F_X(x) \le y] \le y$$

We know that $F_X(x)$, $F_y(y)$ are ~~positive~~ nonnegative & increasing

Consider the interval $(0, x_1)$ and let $y = y_1 > 0$ as indicated above

While $F_X(x) = 0$ on $\forall x \in (0, x_1)$ note $y_1 > 0$ s.t. $F_X(x) < y$

generally, between "jumps" in $F_X(x)$ based on values of X with nonzero probabilities, we will see that $F_X(x) < F_U(y) = y$ - at y_3, etc.

Now consider an $F_X(x)$ value at a "jump", say $X = x_1$ or $X = x_4$

At these "jumps" we find equality, $P[F_X(x) \le y] = y$

Thus based on these two cases $F_y(y) \le F_U(u) = y$

AGRESTI EXERCISE 1.7

1.7 In 20 observations, a new drug is better every time

π: probability that new drug is judged better

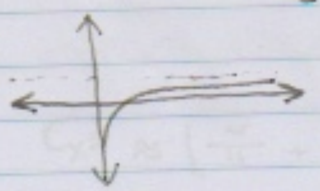
$H_0: \pi = 0.5$ vs. $H_A: \pi \neq 0.5$

(a) Find & sketch the likelihood function. Give the ML estimate of π

Let's use the model $y \sim \text{bin}(n=20, \pi)$

The binomial log likelihood is:

$$L(\pi) = \log\left[\pi^{y}(1-\pi)^{20-y}\right] = y\log(\pi) + (20-y)\log(1-\pi)$$
$$= 20\log(\pi) + (0)\log(1-\pi) = 20\log(\pi)$$



* SKETCH *

We know that the MLE of a binomial dist. is

$\hat{\pi}_{mle} = y/n = 20/20 = 1$ in this case

(b) Conduct a Wald test w/ 95% CI for π. Are results sensible?

Wald test statistic: $z_W = \left[\hat{\pi} - \pi_0\right] / \left[\sqrt{\hat{\pi}(1-\hat{\pi})/n}\right]$

$$= \left[1 - 0.5\right] / \left[\sqrt{(1)(1-1)/20}\right] = \infty$$

Wald 95% C.I.: $\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}} \Leftrightarrow 1 \pm z_{\alpha/2}\sqrt{\frac{1(1-1)}{n}}$

95% CI is 1, a point

No, results are not sensible — this is a point estimate, not a 95% C.I.

AGRESTI 1.7 cont.

c) Conduct a score test, reporting p-value and 95% C.I. Interpret.

$$z_S = \left[\hat{\pi} - \pi_0\right] \Big/ \left[\frac{\sqrt{\pi_0(1-\pi_0)}}{n}\right] = [1-0.5] \Big/ \left[\frac{\sqrt{(0.5)(1-0.5)}}{20}\right]$$

$$= (0.5)/(0.5/20) = 20$$

approximately normal under the null distribution

$z_S \sim N$ under $H_0: \pi = 0.5$

p-value ~~≥ 0.999 with $\alpha = 0.05$~~ < 0.0001

We will reject H_0 at $\alpha = 0.01$

$$C_{X^2} \approx \left[\tilde{\pi} + z_{\alpha/2}\sqrt{\frac{\tilde{\pi}(1-\tilde{\pi})}{n}}\right];\ \tilde{\pi} = \frac{y + (z_{\alpha/2})^2/2}{\tilde{n}};\ \tilde{n} = n + (z_{\alpha/2})^2$$

$\alpha = 0.05 \rightarrow z_{\alpha/2} = z_{0.025} = 1.960$

so $\tilde{n} =$ ~~n~~ $20 + (1.960)^2 = 23.8416$

$\tilde{\pi} = [20 + (1.96)^2/2] / 23.8416 = 0.9194349372$

~~π~~ for: $0.919 \pm (1.96)\sqrt{\frac{(0.919)(1-0.919)}{20}}$

$\Longrightarrow 0.919 \pm 0.1192820102$

$\Rightarrow [0.7997179...898, 1.03828201]$

We are 95% confident that the true value lies between 0.7997... and 1.03828201

d) Conduct a likelihood ratio test & construct 95% CI interpret.

$$L^2 = 2\left[y \log\left(\frac{y}{n\pi_0}\right) + (n-y)\log\left(\frac{n-y}{n-n\pi_0}\right)\right]$$

$$= 2\left[20 \log\left(\frac{20}{20\cdot 0.5}\right) + (20-20)\log\left(\frac{20-20}{20-20\cdot 0.5}\right)\right]$$

$$= 40\log(2) = \underline{12.04119983} \Longrightarrow \text{p-val} < 0.001$$

at $\alpha = 0.01$ reject null hyp. that $\pi = 0.5$

95% C.I. no closed form solution here...
must use SAS?

SAS gives [1,1] as a 95% confidence interval
Again, effectively a point estimate rather than C.I.

c) Conduct an ~~[illegible]~~ exact binomial test & 95% C.I. Interpret

SAS gives 95% CI: $[0.8316, 1.000]$

and a p-value of <0.0001

reject the null that $H_0: \pi = \cancel{0.6}\ 0.5$ at $\alpha = 0.05$

f) Suppose researchers wanted to estimate probability to within 0.05 w/ confidence 95%. If $\pi_T = 0.9$, how large a sample is needed?

let's use Wald test...

$$C_W^2 = \left[\hat{\pi} \pm z_{\alpha/2}\sqrt{\frac{\hat{\pi}(1-\hat{\pi})}{n}}\right]$$

$$z_{0.025}\sqrt{\frac{0.9(1-0.9)}{n}} = 0.05$$

$$(1.960)\sqrt{\frac{0.9(0.1)}{n}} = 0.05$$

$$n = \left(\frac{1.960}{0.05}\right)^2 (0.9)(0.1) = 138.2976 \approx 139$$

A sample of 139 is required.

AGRESTI 1.8

854 green seedings
249 yellow seedings
1103 in total

Test the hypothesis that 3:1 is the true green:yellow ratio. Report the p-value and interpret.

1103 is a sample size large enough to use approximate tests. Let's count "green" as success.

Then $\hat{\pi}_{mle} = y/n = 854/1103 = 0.7742520399$

$H_0: \pi_0 = 0.75$
$H_1: \pi_0 \neq 0.75$

Let's use a score test statistic w/ normal form (not $\chi^2_{df=1}$):

$$Z_S = \frac{\hat{\pi} - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.774 - 0.75}{\sqrt{0.75(0.25)/1103}} = \frac{0.0242520399}{0.0130380571}$$

$$= 1.860096152$$

Under H_0, $S \equiv X \sim AN(0,1)$ 2·pnorm(x, lower.tail = F)

Computation of the associated p-value in R yields:
$2 \cdot Pr(Z \geq 1.80096152) = 0.07170895$

So at the $\alpha = 0.01$ level we fail to reject the null hypothesis that 3:1 is the true green: yellow ratio.

∎