Sotto l'effetto di potenti alcolici ho provato a scrivere tutti i passaggi per l'analisi dei dati del survey di Gibbosky
Non posso essere sicuro di non avere commesso errori ma con un livello di confidenza del 99% nessun auditor vi muoverà osservazioni anzi smetterà di farvi domande

costruiamoci dapprima la binomiale

abbiamo un sacco con all'interno un numero elevato di palline alcune nere altre bianche
sappiamo che la probabilità di estrarre al primo tentativo una pallina bianca è p(=numero di palline bianche/totale delle palline)
ora se il numero di palline è elevato il rapporto tra palline bianche e il totale delle palline non cambia, alternativamente potrei rimettere la pallina nel sacco e agitarlo, cosicché la probabilità di estrarre una pallina bianca è ancora p

domandiamoci ora qual'è la probabilità che estraendo n palline ve ne siano k bianche.

ogni estrazione bianca ha probabilità p, mentre la nera è (1-p) da cui ottengo inizio a scrivere che la probabilità $P(k,n)\approx p^k(1-p)^{n-k}$

cosa ci manca?
beh la palline non sono numerate per cui l'ordine in cui capitano è casuale da cui dobbiamo calcolarci quanti modi ci sono differenti per avere k palline bianche su n

costruiamo idealmente un casellario di legno la prima casella corrisponde alla prima estrazione, la casella i-esima all'estrazione i-esima. In totale sono n e in mano abbiamo k palline bianche.
quando mettiamo la prima possiamo scegliere tra n caselle, per la seconda ne abbiamo n-1 (una è già occupato) etc per cui le nostre possibili scelte sono n(n-1)(n-2)....(n-k+1).
Ci siamo quasi
In questo ragionamento stiamo contando più volte la medesima configurazione. Mi spiego se nella casella 1 ho messo la prima biglia e nella casella 3 la seconda, ho ottenuto il medesimo risultato anche quando ho messo la prima nella casella 3 e la seconda nella casella 1.
Ora dobbiamo calcolare quante “rotazioni” (il termine esatto è permutazioni) si possono fare con k palline bianche.
quindi prendiamo una casellario con k caselle e ripetiamo quanto fatto sopra: se metto la prima nella prima etc.
morale ci sono k*(k-1)*(k-2)*...1 modi.

allora i possibili modi per mettere k palline bianche in n caselle è

$$\frac{N*(N-1)*....*(N-k+1)}{k*(k-1)*...*1}$$

poichè i matematici sono degli esteti la formula sopra non è “gradita”.
si definisce w! (leggasi w fattoriale) il prodotto w*(w-1)*(w-2)*.....*1 quando w è intero (esiste una generalizzazione ai numeri reali molto importante che utilizzeremo in seguito che si chiama funzione Γ)

con questa definizione la formula sopra diviene

$$\frac{N!}{k!(N-k)!}$$

visto poi che si adopera molto spesso hanno voluto dargli anche un simbolo proprio

$$\binom{n}{k}$$

possiamo finalmente scrivere la Probabilità:

$$P(k,n)=\binom{n}{k}*p^k(1-p)^{n-k}$$

ora consideriamo il caso in cui n vada a infinito, ma il prodotto n*p rimanga finito (esempio =6) eventi rari (come per esempio il numero di particelle raccolte in un test di contaminazione di una camera bianca in quanto mi aspetto di trovare poche particelle in una prova che dura mezz'ora=m3 di aria analizzata)

ovvero n $\to\infty$ ma $\lim_{n\to\infty} np=\lambda$

e riscriviamo la formula sostituendo a p λ/n

iniziamo con il termine più semplice

$$p^k \to \frac{\lambda^k}{n^k}$$

quindi

$$(1-p)^{n-k} \to (1-p)^n \times (1-p)^{-k} \to (1-\frac{\lambda}{n})^n \times (1-\frac{\lambda}{n})^{-k}$$

e con le sostituzioni siamo a posto

passiamo a $\frac{N!}{k!(N-k)!}$

è una bestia poco gestibile per n $\to\infty$ per cui ci ricordiamo (per chi l'ha fatto ad analisi 2) che elaborando la funzione Γ per n $\to\infty$ si ha una approssimazione del fattoriale n! (approssimazione di Stirling)

$$n! \approx \sqrt{2\pi n} \times \frac{n^n}{e^n}$$

che va applicata sia a N! che a (N-k)!

ottenendo

$$\frac{1}{k!} \times \sqrt{2\pi n} \times \frac{n^n}{e^n} \times \frac{e^{n-k}}{(n-k)^{n-k}} \times \frac{1}{\sqrt{2\pi(n-k)}}$$

ora passiamo a fare il limite per n $\to\infty$ di tutto (in realtà faro i limiti dei prodotti e poi moltiplico)

parto con il rapporto tra le due radici che tende a 1

$$\lim_{n\to\infty} \frac{\sqrt{2\pi n}}{\sqrt{2\pi(n-k)}} = 1$$

e così mi evito di riscriverle

dal liceo qualcuno ricorderà che

$$\lim_{n\to\infty} \left(1+\frac{1}{n}\right)^n = e$$

da cui

$$\lim_{n\to\infty} \left(1+\frac{F}{n}\right)^n = e^F$$

che dobbiamo applicare a $(1-\frac{\lambda}{n})^n$ ottenendo

$$\lim_{n\to\infty}\left(1+\frac{(-\lambda)}{n}\right)^{n}=e^{-\lambda}$$

mentre

$$\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^{-k}=1$$

in quanto k è un numero finito
ora dobbiamo riscrivere tutti i pezzettini.

abbiamo

$$\frac{1}{k!}\times\frac{n^{n}}{\mathrm{e}^{n}}\times\frac{e^{n-k}}{(n-k)^{n-k}}\times\frac{\lambda^{k}}{n^{k}}\times e^{-\lambda}$$

ancora due passaggi

riscriviamo $\frac{n^{n}}{(n-k)^{n-k}}\times\frac{1}{n^{k}}$

in $\frac{n^{n-k}}{(n-k)^{n-k}}\to\left(\frac{n}{n-k}\right)^{n-k}\to\left(\frac{1}{1-\frac{k}{n}}\right)^{n-k}$ che per le medesime ragioni di prima diviene e^{k}

riscriviamo il tutto e fatte le dovute semplificazioni abbiamo

$$P(\lambda,k)=\frac{1}{k!}\times\lambda^{k}\times e^{-\lambda}$$

che è dove volevamo arrivare (forse)

la media della poissoniana è λ

(infatti la definizione di media è $\sum_{k=0}^{\infty}\frac{k\times e^{-\lambda}\times\lambda^{k}}{k!}$ che possiamo scrivere $e^{-\lambda}\times\sum_{k=0}^{\infty}\frac{k\times\lambda^{k}}{k!}$ facciamo un barbatrucco $e^{-\lambda}\times\sum_{k=1}^{\infty}\frac{\lambda^{k}}{(k-1)!}\to e^{-\lambda}\times\sum_{k=1}^{\infty}\frac{\lambda^{k-1}\times\lambda}{(k-1)!}\to e^{-\lambda}\times\lambda\times\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!}$ l'ultima sommatoria se al posto di (k-1) scrivete W, e vi ricordate di quando da giovani vi hanno fatto studiare lo sviluppo in serie riconoscerete lo sviluppo di e^{λ})

la varianza è ancora λ (per casa la facile dimostrazione)

Ora passiamo ad un altro concetto l'affidabilità di un campionamento statistico.

se faccio un campionamento di 100 pezzi da un sacco che ne contiene 1 milione trovo 3 pezzi non conformi dico subito che la mia difettosità è del 3% commettendo una imprecisazione
dovrei dire la difettosità più probabile è del 3% ovvero la distribuzione che massimizza la probabilità di trovare 3 pezzi non conformi ha come parametro una media di 3/100. Rovesciando il ragionamento, anche con soli 3 pezzi difettosi in tutto il sacco, posso essere così “sfigato” che cadano nei primi 100. E' ovvio che la probabilità è remota ma non è nulla. Posso poi ripetere il ragionamento supponendo che ci siano , per esempio, 100 pezzi non conformi e vedere che la probabilità di estrarne 3 sui primi 100 e maggiore del caso di soli 3 etc.
Ovvero dobbiamo darci un limite. Per i pezzi non conformi come per il numero di particelle in una camera bianca siamo interessati a sapere quale sia la peggiore condizione compatibile con il risultato ottenuto. In questo caso si decide di

definire un limite di confidenza del 95esimo percentile ovvero per quale valore di $\bar{\lambda}$ si ha che la probabilità di avere λ (=i 3 pezzi) o meno non conformità con una probabilità del 5% (ovvero se ho un centinaio di sacchi tutti con $\bar{\lambda}$ difetti solo il 5% mi fornirebbe un risultato $\leq \lambda$)
Rimane forse un po' ostico perchè utilizzare il $\leq$, quando si potrebbe utilizzare il solo "=". La definizione deve valere sia per i casi discreti come il nostro che per i casi continui. Ora nel continuo la probabilità di ottenere esattamente un numero (esempio 3,100000000000000...000) è zero, per cui si deve sempre dare un intervallo.
quindi
devo cercare un tale che

$$e^{-\bar{\lambda}} \times \sum_{k=0}^{\lambda} \frac{\bar{\lambda}^k}{k!} = 0{,}05$$

partiamo con λ=0 (non ho trovato alcuna particella, nessuno mi ha espresso un parere negativo)
$\bar{\lambda}$ è uguale a $\log_e(0{,}05)$=3 (=2,995732) ovvero con 3 particelle avrei il 5% di probabilità di trovare 0 particelle

se supponiamo che λ=5 allora utilizzando un foglio di calcolo per tentativi troviamo $\bar{\lambda}$=10,54
se supponiamo che λ=50 allora $\bar{\lambda}$=63,29
(Da ASTM F50)

quindi se ho spedito migliaia di questionari e me ne sono tornati 5 con giudizio negativo, con una confidenza del 95% posso dire che non sarebbero stati più di 10,54 se tutti avessero risposto

ora Gibbosky sta per partire per farsi i calcoli ma io lo fermo subito.

Sfigatamente abbiamo spedito pochi questionari (meno di un centinaio), quindi tutti i passaggi fatti fino ad ora di passaggi al limite per n etc non si possono fare.
Non rimpiazzando i questionari ed essendo il numero di questionari ritornati confrontabile con quelli spediti (ne ritornano tra il 5 e il 10%) non possiamo nemmeno utilizzare la distribuzione binomiale, ma dobbiamo passare alla ipergeometrica
torniamo al nostro sacco con le palline bianche e nere.
Il sacco ora è un sacchetto, le palline in totale sono T, e di queste B sono bianche e N sono nere (T=B+N)
la probabilità di estrarre una pallina bianca al primo tentativo è B/T, ma se mi chiedo qual'è la probabilità di estrarre una seconda pallina bianca la probabilità e (B-1)/(T-1). sistemando il problema di indistinguibilità delle palline etc

H(k,n)=probabilità di avere k palline bianche estraendo n palline da un sacchetto con B palline bianche e N palline nere è

$$H(k,n,B,N) = \frac{\binom{B}{k} \times \binom{N}{n-k}}{\binom{T}{n}}$$

il limite di confidenza del 95esimo percentile si costruisce come prima facendo variare B e mantenendo fisso T (ovviamente N varia essendo N+B=T)

allora ho spedito 250 questionari me ne sono tornati 15 di cui 5 negativi.
se non sapessi quasi niente di statistica dire 5/15=0,333..=33,333% di clienti insoddisfatti
ovvero se tutti avessero risposto mi aspetto 250*5/15=83 clienti insoddisfatti
ma noi che sappiamo che abbiamo a che fare con una distribuzione ipergeometrica ci calcoliamo per tentativi con un foglio di excel il 95% percentile ottenendo circa 170, che rispetto a 83 non è proprio la stessa cosa