Further Maths Statistics

Expected value = $E(X) = \sum x \, P(X=x)$

$E(X^2) = \sum x^2 \, P(X=x)$

Variance = $Var(X) = E\left((X - E(X))^2\right)$

$= E(X^2) - (E(X))^2$

Conversions = $E(g(x)) = \sum g(x) \, P(X=x)$

$E(aX+b) = a\,E(X) + b$

$E(X+Y) = E(X) + E(Y)$

$Var(aX+b) = a^2 Var(X)$

Poisson : events occuring at a constant rate $X \sim Po(\lambda)$

Events must occur : independently, singly in time or space, at a constant average rate

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!}$$

if $X \sim Po(\lambda)$, $Y \sim Po(\mu)$

then: $X + Y \sim Po(\lambda + \mu)$

$E(X) = \lambda$

$Var(X) = \lambda$

Poisson suitable if mean close to variance

ONLY APPROXIMATION → Binomial approximated to Poisson :

if $X \sim B(n, p)$, then:

$E(X) = np$

$Var(X) = npq$

if n is large, and p is small

$X \approx Po(np)$

Geometric : no. trials to get one success $X \sim Geo(p)$

Events must occur : independently, singly, with a constant probability

$P(X=x) = p(x) = pq^{x-1}$

$E(X) = \mu = 1/p$

$Var(X) = \frac{q}{p^2}$

$P(X \leq x) = 1 - q^x$

~~$P(X > x) = q^x$~~

$P(X \geq x) = q^{x-1}$

Negative Binomial: no. trials to get r successes $X \sim NBin(r, p)$

Events must occur: independently, singly, with constant probability p

$$P(X = x) = {}^{x-1}C_{r-1}\, p^r q^{x-r} \quad \text{or} \quad \binom{x-1}{r-1} p^r q^{x-r}$$

$$E(X) = \frac{r}{p} \qquad Var(X) = \frac{rq}{p^2}$$

Hypothesis testing

1. Define random variables
2. Define distribution and parameters
3. Set out hypotheses in terms of parameter
4. Assume H_0 hypothesis
5. State significance level
6. Compare / calculate critical values / critical region / test statistic
7. Conclusion: relate to H_0 in context of question

- If lost, ~~so~~ show an attempt at as many steps as possible
  - ↳ eg. pretend to try to find hypotheses and crit. region
- 'Actual significance' = probability incorrectly rejecting H_0

Central Limit Theorem

A random sample of size n from any distribution with mean μ and variance σ^2, the sample mean $\bar{X} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$

$\boxed{\text{• no continuity correction}}$

use $\sim$ if $X \sim N(\mu, \sigma^2)$

This bit is probably the hardest

Chi-squared Tests

H_0: There is no difference between the observed and the theoretical distribution
H_1: There is a difference between the observed and the theoretical distribution

Goodness of fit is concerned with measuring how well an observed frequency distribution fits to a known distribution.

- Measure of goodness of fit: $X^2 = \sum \frac{(O_i - E_i)^2}{E_i}$ or $\sum \frac{O_i^2}{E_i} - N$
  (X^2 is better the lower)

- The χ^2 family of distributions can be used to approximate X^2 as long as none of the expected values is below 5. You can use a χ^2 distribution to find the critical region for a measure of goodness of fit

ν = number of cells (columns) after combining − number of constraints

e.g. estimating a parameter

having to sum to a fixed total counts

You then reject H_0 if:
$X^2 > \chi^2_\nu$ (s.g. level)

Contingency Tables

expected frequency $= \frac{\text{row total} \times \text{colum total}}{\text{grand total}}$

$\nu = (h-1)(k-1)$

Probability Generating Functions

If a discrete random variable X has probability mass function $P(X=x)$, then the probability generating function is given by: $G_X(t) = \sum P(X=x)t^x$ where t is a dummy variable

$G_X(1) = 1$ → effectively probabilities always sum to 1

$G_X(t) = E(t^X)$

Distribution of X	$P(X=x)$	P.G.F
Binomial $B(n,p)$	$\binom{n}{x} p^x q^{n-x}$	$(1-p+pt)^n$
Poisson $Po(\lambda)$	$e^{-\lambda} \lambda^x / x!$	$e^{\lambda(t-1)}$
Geometric $Geo(p)$	pq^{x-1}	$\frac{pt}{1-qt}$
Negative B $NBin(r,p)$	$\binom{x-1}{r-1} p^r q^{x-r}$	$\left(\frac{pt}{1-qt}\right)^r$

$E(X) = G'_X(1)$
$Var(X) = G''_X(1) + G'_X(1) - [G'_X(1)]^2$

if $Z = X+Y$, $G_Z(t) = G_X(t) \times G_Y(t)$

if $Y = aX+b$, $G_Y(t) = t^b G_X(t^a)$

~~Quality of Tests~~

$G(0) = P(X=0)$

$G'(0) = 1!\ P(X=1)$

$G''(0) = 2!\ P(X=2)$

$G^{(n)}(0) = n!\ P(X=n)$

$P(X=0) = \frac{G(0)}{0!}$

$P(X=1) = \frac{G'(0)}{1!}$

$P(X=n) = \frac{G^{(n)}(0)}{n!}$

→ One of the consequences of this is that if: $G_X(t) = a + bt + ct^2$

a: $P(X=0)$, b: $P(X=1)$, c: $P(X=2)$

Quality of Tests

Test Conclusion	Actual	Probability
Accept H_0	H_0 True	OK
Reject H_0	H_0 True	P(Type 1) = size = actual significance level
Accept H_0	H_0 False	P(Type II) => True parameter given
Reject H_0	H_0 false	OK : Power = 1 − P(Type II) ← better when higher

To calculate P(Type 1) find the actual significant level. This will just be the probability of it falling in the critical region. (good chance you've already done the calculation)
To calculate P(Type II) find the probability of it falling outside the critical region (prob of being accepted) given the parameter in the question.

Increasing the sample size reduces the ~~[illegible]~~ prob. of Type I and Type II
Increasing the significance level increases P(Type I) and reduces P(Type II)

Power function : 1. Find P(Type II) in terms of the parameter

2. Do 1 − Ans