**#Binomial distribution**

Binomial conditions

1. the trials must be independent

2. the number of trials, n, must be fixed

3. each trial outcome must be classified as a success or a failure

4. the probability of success, p, must be the same for each trial

Binomial distribution:

If p represents probability of success, (1-p) represents probability of failure, n represents number of independent trials, and k represents number of successes.

P(k successes in n trials)=(N Choose K)\*P^k\*(P-1)^k

Among a random sample of 100 employees, how many would you expect to

be engaged at work? Remember: p = 0.13.

> dbinom(8, size = 10, p = 0.13)

Expected value (mean) of binomial distribution: mean=np

􀐖 = 100 x 0.13 = 13

Standard deviation of binomial distribution: sigma=sqrt(np(1-p))

SD =sqrt( 100 x 0.13 x 0.87) = 3.36

Success-failure rule: A binomial distribution with at least 10

expected successes and 10 expected failures closely follows

a normal distribution.

np ≥ 10

n(1-p) ≥ 10

**#Central limit theorem (CTL)**

Central Limit Theorem (CLT): The distribution of sample statistics is nearly

normal, centered at the population mean, and with a standard deviation equal

to the population standard deviation divided by square root of the sample size.

(population mean = mean of the distribution ; SE=(S/sqrt(n))

Conditions for the CLT:

1. Independence: Sampled observations must be independent.

‣ Random sample/assignment

‣ if sampling without replacement, n < 10% of population

2. Sample size/skew: Either the population distribution is normal, or if the population

distribution is skewed, the sample size is large (rule of thumb: n > 30).

CLT for proportions: The distribution of sample proportions is nearly

normal, centered at the population proportion, and with a standard error

inversely proportional to the sample size.

SE=sqrt (P(1-P)/N)

Conditions are same as given above .

Example problem: 90% of all plants species are classified as angiosperms (flowering plants).

Difference between proportions

sqrt { (P1(1-P1)/n1) + (P2(1-P2)/n2) }

Problem: Comparison Handguns course era and US comparison.

**#Backtracking to n for a given ME**

Given a target margin of error, confidence level, and information on the variability of the sample (or the population), we can determine the required

sample size to achieve the desired margin of error.

ME=Z\*(s/sqrt(n))

**#T distribution**

Always centered at 0 (like the standard normal)

‣ has one parameter: degrees of freedom (df) - determines thickness of tails

‣ remember, the normal distribution has two parameters: mean and SD

T=(Obs-null)/SE

pnorm(2, lower.tail = FALSE) \* 2 🡨 normal distribution two sides hypothesis test

pt(2, df = 50, lower.tail = FALSE) \* 2 <- student two sides hypothesis test

**#Power of test**

1. Type 1 error is rejecting H0 when you shouldn’t have, and the probability of doing so is α (significance level).
2. Type 2 error is failing to reject H0 when you should have, and the probability of doing so is β.
3. Power of a test is the probability of correctly rejecting H0, and the probability of doing so is 1 − β

**#anova test**

Problem: vocabulary score and class.

Hypothesis test: If p-value is small (less than α), reject H0.

The data provide convincing evidence that at least one pair of population means are different from each other (but we can’t tell which one).

p-value is large, fail to reject H0.

The data do not provide convincing evidence that at least one pair of population means are different from each other, the observed differences in sample means

are attributable to sampling variability (or chance).

SSE(sum of square errors)=SST(sum squared totals)-SSG(sum squared groups)

Degree of freedom

Dft total=n-1 -- 795 - 1 = 794

Dfg group=k-1-- 4 - 1 = 3

Dfe Error=Dft-Dfg -- 794 - 3 = 791

#Mean squares

MSG = SSG/DfG

MSE = SSE/DfE

F = (MSG/MSE)

pf(21.735, 3, 791, lower.tail = FALSE)

Conditions for ANOVA

1. Independence:

✓ within groups: sampled observations must be independent

✓ between groups: the groups must be independent of each other (non-paired)

2. Approximate normality: distributions should be nearly normal within each group

3. Equal variance: groups should have roughly equal variability

**#small sample proportion**

**##Paul the Octopus predicted 8 World Cup games, and predicted them all**

**##correctly. Does this provide convincing evidence that Paul actually has psychic**

**##powers, i.e. that he does better than just randomly guessing?**

H0: p = 0.5

HA: p > 0.5

1. independence:

2. sample size / skew:

we can assume that his guesses are independent

n = 8 x 0.5 = 4 􀋠 not met

p = 1 distribution of sample proportions cannot be

assumed to be nearly normal

> source("http://bit.ly/dasi\_inference")

> paul = factor(c(rep("yes", 8), rep("no", 0)), levels = c("yes","no"))

> inference(paul, est = "proportion", type = "ht", method = "simulation",success = "yes", null = 0.5, alternative = "greater")

**#chi-square GOF test**

**#chi-square statistic**

**##when dealing with counts and investigating how far the observed**

**##counts are from the expected counts, we use a new test statistic called**

**##the chi-square (χ2) statistic.**

X^2 static = Sigma(1 to k) {(O-E)^2/E)}

O: Observed

E: Expected

K: number of cells

Df=k-1

**Ethnicity white black nat. amer. asian & PI other total**

%in population 80.29% 12.06% 0.79% 2.92% 3.94% 100%

expected # 2007 302 20 73 98 2500

observed # 1920 347 19 84 130 2500

H0: The observed counts of jurors from various race/ethnicities

follow the same ethnicity distribution in the population.

HA: The observed counts of jurors from various ethnicities do not

follow the same race/ethnicity distribution in the population.

X^2= (1920 - 2007)^2/2007+ (347 - 302)^2/302 + (19 - 20)^2/20+ (84 - 73)^2/73+ (130 - 98)^2/98

= 22.63

df = k - 1 = 5 - 1 = 4

> pchisq(22.63, 4, lower.tail = FALSE)

[1] 0.0002

**# chi-square independence test**

**# two categorical variables, at least 1 with >2 levels**

X^2: same as above

Df: (R-1)\*(C-1)(R: Number of rows, C: number of columns)

**#Conditions for the chi-square test:**

**#1. Independence: Sampled observations must be independent.**

**#‣ random sample/assignment**

**#‣ if sampling without replacement, n < 10% of population**

**#‣ each case only contributes to one cell in the table**

**#2. Sample size: Each particular scenario (i.e. cell) must have at least 5 expected cases.**

obese 81 103 147 331

not obese 59 326 277 962

total 440 429 424 1293

What is the overall obesity rate

in the sample?

331 / 1293 = 0.256

If in fact weight and relationship status are independent (i.e. if in fact H0 is true) how

many of the dating people would we expect to be obese? How many of the cohabiting

and married?