Central limit thm:

Political view	Population
1	5%
2	10%
3	15%
4	20%
5	20%
6	10%
7	10%
8	5%
9	3%
10	2%
	100%

any distribution

true mean $= 1(0.05) + 2(0.01) + \dots + 10(0.02)$

$= 4.67$

Sample 1000 times, 100 people each

$[\mu_1, \mu_2 \dots \mu_{1000}]$ → normally dist around mean of 4.67

law of large numbers

Political view	Population
1	5%
2	10%
3	15%
4	20%
5	20%
6	10%
7	10%
8	5%
9	3%
10	2%
	100%

true mean $= 1(0.05) + 2(0.01) + \dots + 10(0.02)$

$= 4.67$

Sample 10 people, mean $\neq 4.67$

Sample 100 people, mean $\rightarrow 4.67$

Sample 1000 people mean $\simeq 4.67$

Sample entire people mean $= 4.67$

$$\text{Sample size } n = \left(\frac{z\,\sigma}{E}\right)^2$$

$z \rightarrow z$ score of confidence level
1.96 for 95%

$\sigma \rightarrow$ sd of population

$E \rightarrow$ margin of error desired

Measures of Central tendency

Mean(x) $\mu = E[X] = \int_{-\infty}^{\infty} x f(x)\, dx$

Var(x) $\text{var}(x) = E(x-\mu)^2 = E(x-E[x])^2 = E(x^2) - (Ex)^2$

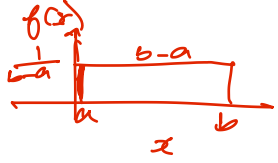
$\sigma = \sqrt{\text{var}(x)}$

$E(X/x=y) = \int_{-\infty}^{\infty} x f_{x/y}(x/y)\, dx$

$\text{cov}(x,y) = E((x-Ex)(y-Ey)) = E(xy) - (Ex)(Ey)$

$\rho(x,y) = \frac{\text{cov}(x,y)}{\sigma_x \sigma_y}$

Uniform Dist



$$f(x) = \frac{1}{b-a}$$

$$E[x] = \int_a^b x\, f(x)\, dx$$

$$= \int_a^b x \cdot \frac{1}{b-a}\, dx = \frac{1}{b-a} \left. \frac{x^2}{2} \right|_a^b = \frac{b^2-a^2}{2(b-a)}$$

$$= \frac{b+a}{2}$$

Hypothesis Testing

H_0 :

H_a : reject or not reject

test statistic ✓ (z, t, Chi-sq) — p-value ✓ — p-value vs α ✓

one tailed: $H_0: \mu = \mu_0$

$H_a: \mu < \mu_0$ or $\mu > \mu_0$

more powerful - directional

in campaign more effective? have higher voter turnout.

two tailed: $H_0: \mu = \mu_0$

$H_a: \mu \neq \mu_0$

non-directional

new policy significant effect on unemployment

Z-test

assumed test statistic follows normal distribution under null hypothesis.

Sample size is large ($30+$ for CLT)

σ is known (pop variance)

$$\text{Z-statistic} \quad z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \sim N(0,1)$$

$\bar{x}$ → sample mean
μ_0 → pop mean
σ → pop sd
n → sample size

e.g.: significant diff btw means of 2 groups or mean differ significantly from known value

---

campaign effectiveness:

nos. of contacts voted voted for BFK $\mu = 0.40$

sd of past elections results $= 0.05 = \sigma$

New strategy (sample 1000 voters)

$\times$ H_0: New strategy doesn't ↑ vote support (μ_0) $\bar{x} = \mu$

✓ H_a: ↑ $\bar{x} > \mu$

1000 voters ≡ 45% support $= \bar{x}$

$$z = \frac{0.45 - 0.40}{0.05/\sqrt{1000}} = \frac{0.05}{0.05} \times \sqrt{1000} = 31.62$$

$31.62 > 1.645$ (1-tailed test, 5% sign level)

reject null hypo

→ Prob of obsv 5% large increase by random chance is low

1.645 → 95% of area under normal curve lies to the left of 1.645

t-test

(1) sample size is small ($n < 30$)
pop var is unknown (σ)

t-statistic $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$

where $s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$

1) indp/unpaired t-test
compare means of 2 indp groups
(like test scores)

t-dist similar to normal distribution
but larger tails

2) paired / dep t-test
same group, diff times.
(test scores before & after extra classes)

example

indp: two should be similar.
indp samples

in their age diff btw Dem & Rep

indep t-test

Dem Sample	Rep Sample
30 voters	30 voters
45 age mean	50 yr mean
5 sd age	6 age sd

H_0: No diff
H_1: there is diff

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{-5}{\sqrt{11/30}} = -3.5$$

$df = n_1 + n_2 - 2$

p-value = 0.00077 < 0.05 → reject H_0

(how is p calculated)
t-distribution (~~used~~ pop sd is unknown)
$n \downarrow$ ⇒ sample dist of mean ~~of~~ does not follow normal but t-dist

= area under t-curve that is more extreme than -3.5

2-tailed: both tails

→ prob of obtain a value more extreme than -3.5 if H_0 is true

Chi-squared test (~~tests goodness of fit~~)

Determine if there is a significant association btw categorical variables

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

$O_i \Rightarrow$ observed value

$E_i \Rightarrow$ expected value

df = # categories in distribution

categorical vars are indep

	E_1	E_2
Y_1	10	20
Y_2	15	30

~~dpnt~~

e.g.

age	Dem	Repub	
Y	100	150	250
O	180	120	300
	280	270	550

H_0: No association

H_a: Yes association

$$E_{ij} = \frac{\text{Row total}_i \times \text{col total}_j}{\text{Grand total}} =$$

$\frac{250 \cdot 280}{550}$	$\frac{250 \cdot 270}{550}$
$\frac{180 \cdot 300}{550}$	$\frac{120 \cdot 300}{550}$

$=$

127.27	122.73
152.73	147.27

$$\chi^2 = \frac{(100-127.27)^2}{127.27} + \frac{(150-122.73)^2}{122.73} + \frac{(180-152.73)^2}{152.73} + \frac{(120-147.27)^2}{147.27}$$

$$= 21.03$$

$$DF = (n_{row}-1)(n_{col}-1) = (2-1)(2-1) = 1$$

p-value = 0.0000045 < 0.05

Reject the null hypothesis

chi-square dist

p-value: If the null hypo is true, what is the prob of observing value this extreme (or more extreme)

2-tail test

55% with part A

new policy affects or not

H_0: unchanged at 55%

H_a: changes from 55%

2-tailed test

$n = 1000$ voters

520 supported part A = 0.52

480 not supported part A = 0.48

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)}/\sqrt{n}} = \frac{0.52 - 0.55}{\sqrt{0.52(0.48)}} \times \sqrt{1000}$$

$$= -1.91$$

$$-1.96 < -1.91 < 1.96$$

within rom

can't reject null hypo

2 tail

assumption

pop : normal dist

esp for sample size < 30

2 ~~fat~~

Distributions

Normal (Gaussian)
Uniform
Binomial
Poisson
Exponential
chi-sq
t-dist
F-dist

Normal

Histogram, QQ plot
Shapiro-Wilk test
Kolmogorov-Smirnov test

pdf

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Symmetric, mean = median = mode

Binomial

Uniform

$$f(x) = \frac{1}{b-a}$$

t-test example 2 (dep/paired)

Same ppl. diff time (before + after policy)

50 ppl, before = mean 6, sd 1.5

after = mean 7, sd 1.2

Ho: no effect.

Ha: significant effect

$$t\text{-stat} = \frac{\bar{d}}{sd/\sqrt{n}} = \frac{|7-6|}{2.06/\sqrt{50}} = 2.63$$

p-value = 0.0114 < 0.05 (statistically significant)

Do not reject Ho