Recall :-

Hypothesis Test

z-test :- Testing for sample mean when σ is known.

$H_0 : \mu = c$ (Null)

$H_A : \begin{cases} \mu < c \\ \mu > c \\ \mu \neq c \end{cases}$ = alternative

Sample: $X_1, X_2, \ldots, X_n$ from population

Compute: $\dfrac{\sqrt{n}(\bar{X} - c)}{\sigma}$ $\bar{X}$ = mean

Fix $\alpha \in (0,1)$ and Find $z_{\alpha/2}$: $\mathbb{P}(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$

$Z \sim \text{Normal}(0,1)$

<u>check:</u> $\frac{\sqrt{n}(\bar{x}-c)}{\sigma} > z_{\alpha/2}$

if it happens then we would reject the null hypothesis.

$\Leftrightarrow$ Reject the null hypothesis if

$$\mathbb{P}\left(Z \geq \frac{\sqrt{n}(\bar{x}-c)}{\sigma}\right) < \alpha$$

<u>t-test:-</u> Test sample mean when σ is not known.

$H_0: \mu = c$ (Null) $\qquad H_A: \begin{cases} \mu < c \\ \mu > c \\ \mu \neq c \end{cases}$ = alternative

Sample: $x_1, x_2, \ldots, x_n$ from population

Compute: $\dfrac{\sqrt{n}(\bar{x} - c)}{S}$ $\qquad$ S - sample variance

$\bar{x}$ - mean

Fix $\alpha \in (0,1)$

Reject null hypothesis if

$$\mathbb{P}\left(T > \frac{\sqrt{n}(\bar{x} - c)}{S}\right) < \alpha \qquad T \sim t_{n-1}$$

- Likelihood Ratio test statistic and derived the above test [last week notes]

Let $n \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Bernoulli (p) random variables.

z-test :-

Assume convergence is "good" that

Normality assumption holds

We want to test:

Null Hypothesis : $p = 0.5$

Alternative Hypothesis: $p \neq 0.5$

Use Binomial Central Limit Theorem that

$$\frac{\sqrt{n}(\bar{X} - p)}{\sqrt{p(1-p)}} \xrightarrow{d} Z,$$

Apply z-test

where Z is standard Normal.

Hypothesis Testing– Proportions

in built test

Use `prop.test`. Suppose $n = 100$, $\bar{X} = 0.43$.

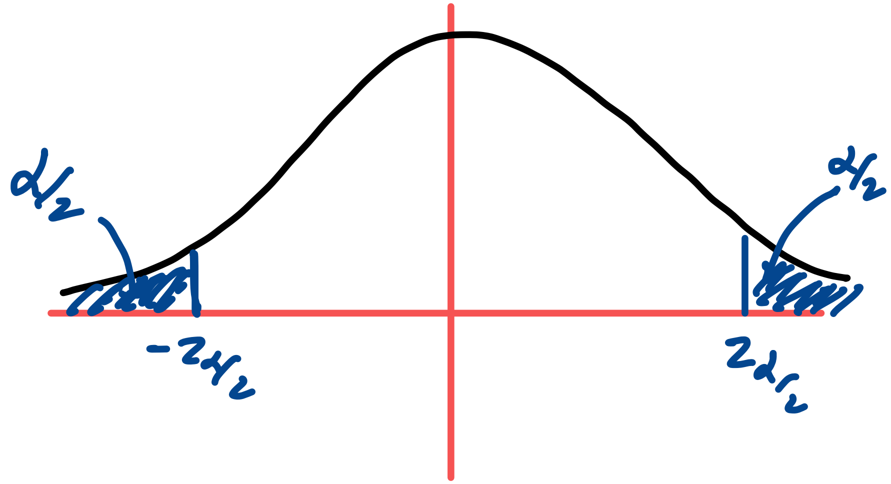
```
> prop.test(43,100)

        1-sample proportions test with continuity correction

data:  43 out of 100, null probability 0.5
X-squared = 1.69, df = 1, p-value = 0.1936
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
 0.3326536 0.5327873
sample estimates:
   p
0.43
```

Hypothesis Testing– Proportions

`prop.test` does the following:



- Computes $P(\mid Z - 0.5 \mid \geq \mid \frac{\sqrt{n}(\bar{X}-0.5)}{0.5} - 0.5 \mid)$ towards p-value.

- Finds $100(1-\alpha)\%$- Confidence Interval by finding the region of p where

$$\mid \frac{\sqrt{n}(\bar{X}-p)}{\sqrt{p(1-p)}} \mid < z_{\frac{\alpha}{2}},$$

where $P(Z > z_{\frac{\alpha}{2}}) = \frac{\alpha}{2}$.

Hypothesis Testing: z-test

Ex. Write a code for t-test

$$\mathbb{P}(|Z| \le z) = \alpha$$

$$\mathbb{P}\left(Z \geq \frac{\sqrt{n}(\bar{X}-\mu)}{\sigma}\right)$$

The below is code for z test and Confidence interval for a data x.

```
> ztestci = function(x, mu=0, sigma=1, alpha=0.95){
+ z = qnorm( (1-alpha)/2, lower.tail=FALSE)
+ sdx = sigma/sqrt(length(x))
+ pvalue = pnorm((mean(x) - mu)/sdx,
+ lower.tail=FALSE)
+ c(mean(x) - z*sdx, mean(x) + z*sdx,  pvalue)
+ }
> x=c(75,76,73,75,74,73,76,73,79) ;y = ztestci(x,76,1.5)
> y
[1] 73.9089069 75.8688709  0.9868659
```

$H_0: \mu = 76$

$H_A: \mu > 76$

$\bar{x} + z\frac{\sigma}{\sqrt{n}}$

$\bar{X} - z \cdot \frac{\sigma}{\sqrt{n}}$

Hypothesis Testing: *t*-test

```
> t.test(x, mu=74)

        One Sample t-test

data:  x
t = 1.3571, df = 8, p-value = 0.2118
alternative hypothesis: true mean is not equal to 74
95 percent confidence interval:
 73.37848 76.39930
sample estimates:
mean of x
 74.88889
```

Hypothesis Testing: *t*-test

```
> wilcox.test(x,mu=74,alt="greater")

        Wilcoxon signed rank test with continuity correction

data:  x
V = 27, p-value = 0.1108
alternative hypothesis: true location is greater than 74
```

Application: one needs to compare two population

Test for equality of means when variance is known

Assume: $X \sim \text{Normal}(\mu_1, \sigma_1^2)$

$Y \sim \text{Normal}(\mu_2, \sigma_2^2)$

— σ_1 is known and σ_2 is known

$H_0: \mu_1 = \mu_2$ vs $H_A: \mu_1 \neq \mu_2$

$\mu_1 - \mu_2 = 0 \rightsquigarrow$ **check:** $\bar{X} - \bar{Y} \sim$ close to 0

Sample :- $X_1, X_2, \ldots, X_{n_1}$ from X

$Y_1, Y_2, \ldots, Y_{n_2}$ from Y

Under our assumptions :

$$\bar{X} - \bar{Y} \sim \text{Normal}\left(\mu_1 - \mu_2 \,,\, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)$$

Test :- $Z \sim \text{Normal}(0,1)$

Fix $\alpha \in (0,1)$

If $$\mathbb{P}\left(|Z| \geq \frac{|\bar{X} - \bar{Y}|}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right) < \alpha$$

Then reject null hypothesis.

Test for proportions when variance is not known

Assume: "Two coins" — p_1 – Prob of heads of coin 1

p_2 – Prob of heads of coin 2

$H_0: p_1 = p_2$

Sample:- $X_1^{(1)}, \ldots, X_n^{(1)} \dashrightarrow \hat{p}_1 = \bar{X}^{(1)}$

$X_1^{(2)}, \ldots, X_n^{(2)} \dashrightarrow \hat{p}_2 = \bar{X}^{(2)}$

Statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\ldots}}$$

"Pooled variance"

$$\hat{p} = \frac{\hat{p}_1 + \hat{p}_2}{2}$$

$$\frac{\hat{p}(1-\hat{p})\,2}{n}$$

Use $Z \sim \text{Normal}(0,1)$ $\equiv$ Requires proof

Hypothesis Testing: Two Sample Proportion-test

$X^{(1)}_1, \ldots, X^{(2)}_{n_1}$ ... from population 1

$X^{(2)}_1, \ldots, X^{(2)}_{n_2}$ ... from population 2

- Want to test if proportion of success $p_1 = p_2$ between two populations.
- Let $\hat{p}_1 = \bar{X}^{(1)}$ and $\hat{p}_2 = \bar{X}^{(2)}$
- The statistic is

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1} + \frac{1}{n_2})}},$$

$$\hat{p} = \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}$$

Large n_1, n_2 assume normality for Z

Requires proof

Assume: value of the variance(s) is not known but they are equal

Let $n, m \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ_X,σ_1^2) and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. Normal(μ_Y,σ_2^2).

Requires a proof

Test Statistic:

$$T := \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}\sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t_{n+m-1}$$

- Equal Variance: $\sigma_1 = \sigma_2$

$$S^2_{pooled} = \frac{(n-1)S_X^2 + (m-1)S_Y^2}{n+m-2}$$

$$H_0: \mu_x = \mu_y \quad \text{vs} \quad H_A = \begin{cases} \mu_x > \mu_y \\ \mu_x < \mu_y \\ \mu_x \neq \mu_y \end{cases}$$

Assume: value of the variance(s) is not known but they are unequal

Let $n, m \geq 1$, $X_1, X_2, \ldots, X_n$ be i.i.d. Normal(μ_X,σ_1^2) and $Y_1, Y_2, \ldots, Y_m$ be i.i.d. Normal(μ_Y,σ_2^2).

Test Statistic: $\tilde{T} := \dfrac{\bar{X} - \bar{Y} - (\mu_x - \mu_Y)}{S_{pooled}}$ $\quad \tilde{T} \sim t_d$

- Equal Variance: $\sigma_1 \neq \sigma_2$

$$S^2_{pooled} = \frac{S^2_X}{n-1} + \frac{S^2_Y}{m-1}$$

$d = \ldots$ complicated expression

χ^2 - Goodness of fit test :-

G. Mendel :- (pea)

- seed shape a and A governed by allele
- cross breed to produce

aa, aA, AA

Estimate :- $\mathbb{P}(aa) = ..$ $\mathbb{P}(aA) = ..$ $\mathbb{P}(AA) = ..$

$\underbrace{\text{Genetic laws}}_{\text{Hypothesis}}$ $\xleftarrow[\text{??}]{\text{Verify}}$ observations from cross-breeding

R. A. Fisher :- "Controversy" data was not repeatable. — 1936 Annals of statistics

" Data was too good a fit for the distribution".

Based on test :- identify if the data comes from a distribution

χ^2- goodness of fit test

Some questions:

Q1 • Are the dice we roll in our experiments in class really fair ? were the die really fair?

Q2 • Are two populations X and Y actually independent ?

Rephrase:

- How well the distribution of the data fit the model ?
- Does one variable affect the distribution of the other ?

Specific Question:

- To understand how "close" are the observed values to those which would be expected under the fitted model ?

Towards Answer:

- In this case we seek to determine whether the distribution of results in a sample could plausibly have come from a distribution specified by a null hypothesis.
- The test statistic is calculated by comparing the observed count of data points within specified categories relative to the expected number of results in those categories (under Null).

$x_1, \ldots, x_n$ – sample from X

$H_0: X \sim \begin{cases} \text{Normal} \\ \vdots \\ \text{Bernoulli}(p) \end{cases}$

$\boxed{\text{Test ?}}$

χ^2- goodness of fit test

- Let T be a random variable with finite range $\{c_1, c_2, \ldots, c_k\}$ for which

Null Hypothesis $\quad P(T = c_j) = p_j > 0$ for $1 \leq j \leq k$.

Suppose there are k possible outcomes and each occurring with a specified Probability.

- Let $X_1, X_2, \ldots, X_n$ be the sample from the distribution T and let

Typo: Y_j = |{ k : X_k =c_j}| $\quad Y_j = |\{j : X_j = c_j\}|$ for $1 \leq j \leq k$..

"Counting the number of sample points in each bin"

Y_j is the number of sample points whose outcome was c_j

- Then the statistic

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

Pearson's Chi-square Test Statistic

χ^2- goodness of fit test

$$\mathbf{X}^2 := \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j} \equiv \sum_{j=1}^{k} \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}}$$

- $\mathbf{X}^2$- has χ^2_{k-1} degrees of freedom, assymptotically as $n \to \infty$. Requires a proof which we will omit for this course
- Null Hypothesis: Distribution comes from Multinomial with parameters $p_1, p_2, \ldots, p_k$
- Alternate Hypothesis: Distribution comes from Multinomial with parameters with at least one parameter different from $p_1, p_2, \ldots, p_k$

Fix level of significance “alpha”

And use the distribution fact about X^2 — as chi-square to compute the p-value

Example has three outcomes: NDA, UPA, Third-Front

Probability of each outcome: 0.38, 0.32,0.3

Observed: 35, 40, 25

Sample Size n = 100

Example:

We divide the political parties in India into 3 large alliances: NDA, UPA, and Third-Front. In the previous election the support had been 38%, 32% and 30% support respectively. Super-Nation TV channel takes a sample of 100 people and finds that there are 35 for NDA, 40 for UPA and 25 for Third-Front. It concludes that the vote share has not changed. Is this hypothesis correct ?

Expected :== (38, 32, 30)

Observed - Expected : === (35-38, 40-32, 25-30)

Contigency Tables

- Bivariate Data is often presented as a two-way table.

- For example in Dengue Data from Manipal Hospital

```
> y = read.table("dengueb.csv", header=TRUE)
 > head(y)              > tail(y)
   DIAGNO BICARB1          DIAGNO BICARB1
 1    DSS    16.2      45       D    22.0
 2    DSS    22.0      46       D    16.6
 3    DSS    16.0      47       D    18.3
 4    DSS    21.3      48       D    23.0
 5    DSS    19.0      49       D    24.0
 6    DSS    18.7      50       D    21.0
```

Contigency Tables

- Bivariate Data is often presented as a two-way table.
- For example in Dengue Data from Manipal Hospital

```
           Diagnosis
Cat.Marker  D DSS
         0  0   6
         1 17  15
         2  8   4
```

where we have grouped values of Marker to be $0, 1, 2$ depending on the values being less than or equal to 16, between 16 and 21, and greater than 21.

χ^2- test of independence

Specific question:

- Does one variable affect the distribution of the other ?

Notation:

- Let n_r be the number of rows in the table.
- Let n_c be the number of columns in the table.
- Let $n = n_r n_c$ be the total number of observations.

If marker does not work then the diagnosis should be independent of the marker.

Model:

- Let $T \equiv (p_{ij})$ with $1 \le i \le n_r, 1 \le j \le n_c$ be a probability distribution on $\{(i, j) : 1 \le i \le n_r$ and $1 \le j \le n_c\}$
- Let $p_i^R = \sum_{j=1}^{n_c} p_{ij}$ and $p_j^C = \sum_{i=1}^{n_r} p_{ij}$

χ^2- test of independence

- Null Hypothesis: Variables are independent i.e

$$p_{ij} = p_i^R p_j^C \text{ for all } 1 \leq i \leq n_r \text{ and } 1 \leq j \leq n_c$$

- Alternate Hypothesis: Variables are not independent

χ^2- test of independence

- Let y_{ij} record the frequency in the (i, j) cell.
- Let

$$\hat{p}_i^R = \frac{\sum_{j=1}^{n_c} y_{ij}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} y_{ij}} \text{ and } \hat{p}_j^C = \frac{\sum_{i=1}^{n_r} y_{ij}}{\sum_{i=1}^{n_r} \sum_{j=1}^{n_c} y_{ij}}$$

Individual Probabilities

Let

$$\hat{p}_{ij} = \hat{p}_i^R \hat{p}_j^C$$

Under Independence

and

$$\mathbf{X}^2 := \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

χ^2- test of independence

- Test Statistic:

Omit Proof for this class

$$\mathbf{X}^2 := \sum_{i=1}^{n_r} \sum_{j=1}^{n_c} \frac{(y_{ij} - n\hat{p}_{ij})^2}{n\hat{p}_{ij}}$$

  is χ^2_q distributed assymptotically as $n \to \infty$ with $q = (n_r - 1)(n_c - 1)$ degrees of freedom.
- Decide on level of significance: α
- Compute p-value:

$$\mathbb{P}(\chi^2_q \geq X^2)$$

- Reject Null Hypotheis:

  if p-value is less than α

χ^2- test of independence

For example in Dengue Data from Manipal Hospital:

```
> T = table(Cat.Marker, Diagnosis)
> T
          Diagnosis
Cat.Marker  D DSS
         0  0   6
         1 17  15
         2  8   4
```

Can we test if the Marker value is independent of the characterisation of Dengue as normal or severe ?

Doctor's needs:

A patient arrives with Dengue

Based on Marker doctor needs to decide on Treatment

Statistical test performed:

We collected data of patients : Marker and final diagnosis

We test if Marker is independent of Diagnosis

χ^2- test of independence

For example in Dengue Data from Manipal Hospital:

```
> chisq.test(T)

        Pearson's Chi-squared test

data:  T
X-squared = 7.4583, df = 2, p-value = 0.02401
```