Test for Sample Mean when σ is known

$X_1, X_2, \ldots, X_n$ are i.i.d. $X \sim \text{Normal}(\mu, \sigma)$,

where $\mu \in \mathcal{P} = \mathbb{R}$ is unknown and $\sigma > 0$ is known.

Test the Null hypothesis that $\mu = a$,

or equivalently, that $\mu \in \mathcal{P}_0 = \{a\}$.

- Assume w.l.o.g $\sigma = 1$ (Ex: standardise otherwise)
- Towards a Test-statistic: $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$; Estimate of μ

"Test": If $\bar{X}$ "far away" from a then we reject the null-hypothesis.

If $\mu = a$; $X \sim N(a,1)$ $\bar{X} \overset{d}{=} \text{Normal}\left(a, \frac{1}{n}\right)$

$\Rightarrow T(X_1, \ldots, X_n) := \sqrt{n}(\bar{X} - a) \sim N(0,1)$

Then how to judge "far away".

Take $Y_1, \ldots, Y_n \overset{d}{=} N(a,1)$ independent

$T(Y_1 \ldots Y_n) = \sqrt{n}(\bar{Y} - a) \sim N(0,1)$

- Probabilities of observing a value of $T(X_1, \ldots, X_n)$ by $T(Y_1, \ldots, Y_n)$

will help us judge if

$\bar{x}$ is "far away" from a

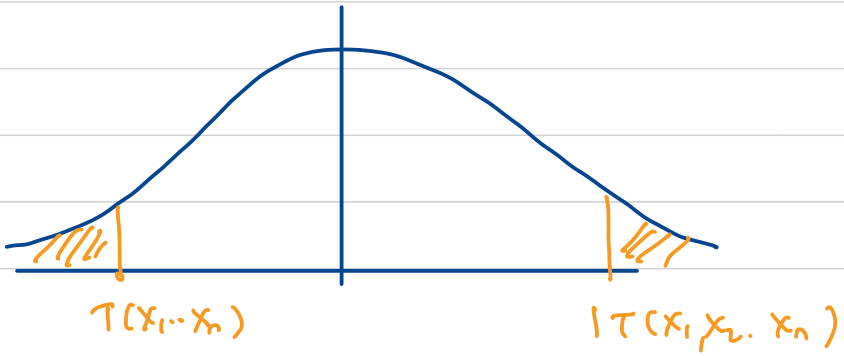
- Symmetry in distribution of Y_i around a

$\therefore$ if $x_1, \dots, x_n$ did not arise from $N(a,1)$ (i.e not have the same distribution as $Y_1 \dots, Y_n$

then

(*) $\mathbb{P}(|T(Y_1 \dots, Y_n)| > |T(x_1 \dots x_n)|)$

Should be small.



Test :- (two sided)

α :- Fix a level of significance α.
say = 5% = 0.05

Test statistic :- Compute $T(X_1, \ldots, X_n) = \sqrt{n}(\bar{x} - a)$

p-value :- let $Z \overset{d}{=} N(0,1)$ find

the $\underbrace{P(|Z| > |\sqrt{n}(\bar{x} - a)|)}_{\text{p-value}}$

If p-value $< \alpha$ then we reject the null hypothesis.

Test for Sample Mean when σ is known

$X_1, X_2, \ldots, X_n$ are i.i.d. $X \sim \text{Normal}(\mu, \sigma)$,

where $\mu \in \mathcal{P} = \mathbb{R}$ is unknown and $\sigma > 0$ is known.

Test the Null hypothesis that $\mu \leq a$,

or equivalently, that $\mu \in \mathcal{P}_0 = \{(-\infty, a]\}$.

Assume w.l.o.g $\sigma = 1$ (Ex: standardise otherwise)

Towards a Test-statistic: $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$; Estimate of μ

"Test": If $\bar{X}$ **is much larger** than a (one sided) then we reject the null-hypothesis

α :- Fix a level of significance α.
say $= 5\% = 0.05$

Test statistic :- Compute $T(X_1, \ldots, X_n) = \sqrt{n}(\bar{x} - a)$

p-value :- let $Z \overset{d}{=} N(0,1)$ find

the $\underbrace{P(Z > \sqrt{n}(\bar{x} - a))}_{\text{p-value}}$

If p-value $< \alpha$ then reject null hypothesis

Test for Sample Mean when σ is unknown

$X_1, X_2, \ldots, X_n$ are i.i.d. $X \sim \text{Normal}(\mu, \sigma)$,

where $\mu \in \mathcal{P} = \mathbb{R}$ and $\sigma > 0$ are unknown.

Test the Null hypothesis that $\mu = a$,

or equivalently, that $\mu \in \mathcal{P}_0 = \{a\}$.

Towards a Test-statistic : $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$; Estimate of μ

"Test" : If $\bar{X}$ "far away" from a then we reject the null-hypothesis.

Issue: $\bar{X} \overset{d}{=} N\left(a, \frac{\sigma^2}{n}\right)$ σ is unknown

$T(X_1 \ldots X_n) = \sqrt{n}(\bar{X} - a) \sim N(0, \sigma^2)$

is not parameter free!

$$S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$$

Set $T(X_1 \ldots X_n) = \frac{\sqrt{n}(\bar{X} - a)}{S} \sim t_{n-1}$

↑ if null was true

Test:

$X_1, X_2, \ldots, X_n$ are i.i.d. $X \sim \text{Normal}(\mu, \sigma)$,

where $\mu \in \mathcal{P} = \mathbb{R}$ and $\sigma > 0$ are unknown.

Test the Null hypothesis that $\mu = a$,

or equivalently, that $\mu \in \mathcal{P}_0 = \{a\}$.

α :- Fix a level of significance α.

say $= 5\% = 0.05$

Test statistic :- $\bar{X} = \frac{X_1 + \cdots + X_n}{n}$ $\quad S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

$$T(X_1 \ldots X_n) = \frac{\sqrt{n}\,(\bar{X} - a)}{S}$$

p-value :- let $T \overset{d}{=} t_{n-1}$ find

the $P(|T| > |T(X_1 \ldots X_n)|)$

p-value

If p-value $< \alpha$ then reject null hypothesis

χ^2- goodness of fit test

Some questions:

1. Are the dice we roll in our experiments in class really fair ?

   Rephrase:

   1.1 How well the distribution of the data fit the model ?

   1.2 Does one variable affect the distribution of the other ?

The parameter space for the problem is:

$$\mathcal{P} = \{p = (p_1, p_2, \ldots, p_6) : 0 \leq p_i \leq 1 \text{ for all } i, \sum_{i=1}^{6} p_i = 1\}$$

Question :-

① How close are the observed values to those which we could expect under the fitted model ?

Null hypothesis :-

$P_0 = \{ (1/6, 1/6, 1/6, 1/6, 1/6, 1/6,) \}$

Test-statistic :-

Could Compare " observed " & " Expected "

χ^2- goodness of fit test (General – Set up)

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

$$P(T = c_j) = p_j > 0 \text{ for } 1 \leq j \leq k.$$

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

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

Y_j is the number of sample points whose outcome was c_j

3. Then the statistic

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

- Theorem :- $X^2 \sim \chi^2_{k-1}$ distribution as $n \to \infty$.

- Null hypothesis : $X_1, \ldots, X_n$ are from multinomial $(p_1, \ldots, p_k)$

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

1. Large values of $\mathbf{X}^2$ indicate that the observed counts don't match expected counts.
2. Large values of $\mathbf{X}^2$ indicates evidence that Null is not correct.

Test :-

α - fix level of significance

say 5% $\equiv$ 0.05

Test Statistic :- Compute

$$X^2 = \sum_{j=1}^{k} \frac{(Y_j - np_j)^2}{np_j}$$

P-value: χ^2_{k-1} — r.v. with $k-1$ d.o.f.

Find $\mathbb{P}(\chi^2_{k-1} \geq X^2)$

If p-value $< \alpha$ then reject the null hypothesis

χ^2- goodness of fit test

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 ?

Null : $p_1 = 0.38$, $p_2 = 0.32$, $p_3 = 0.3$

χ^2- goodness of fit test

1. Null Hypothesis: Vote Share is $(38, 32, 30)$
2. Level of Significance: 0.05
3. Data: Sample Vote share is $(35, 40, 25)$

```
> Xsquared = sum(((x-n*prob)^2)/(n*prob))
> Xsquared
[1] 3.070175
> pchisq(Xsquared, df = 3 -1, lower.tail=FALSE)
[1] 0.2154368
```

$P(\chi^2_2 \geq X\text{squared})$

Since p-value is not smaller than 0.05 we do not reject the null hypothesis.

χ^2- goodness of fit test

Summary

1. Test Statistic:

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

2. Decide on level of significance: α

3. Compute p-value:

$$\mathbb{P}(\chi^2_{k-1} \geq X^2)$$

4. Reject Null Hypotheis:

if p-value is less than α

χ^2- test of independence

Specific question:

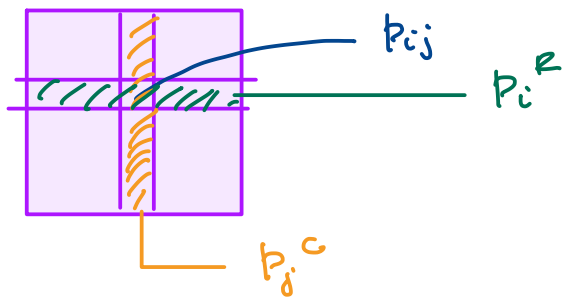
1. Does one variable affect the distribution of the other ?

Notation:

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

Independence

$$p_{ij} = p_i^R \, p_j^C$$



Model:

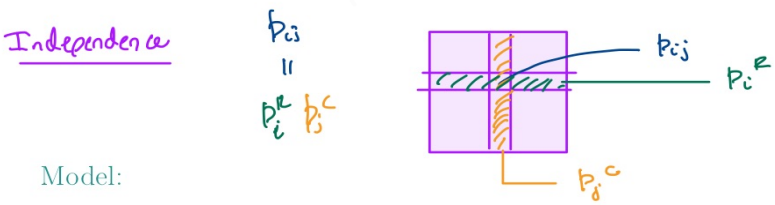
1. Let $T \equiv (p_{ij})$ with $1 \leq i \leq n_r, 1 \leq j \leq n_c$ be a probability distribution on $\{(i,j) : 1 \leq i \leq n_r \text{ and } 1 \leq j \leq n_c\}$
2. Let $p_i^R = \sum_{j=1}^{n_c} p_{ij}$ and $p_j^C = \sum_{i=1}^{n_r} p_{ij}$

χ^2- test of independence

Specific question:

Notation:

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



Model:

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

Framework

1. Null Hypothesis: Variables are independent i.e

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

2. Alternate Hypothesis: Variables are not independent

χ^2- test of independence

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

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

Let

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

and

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

is χ^2_q distributed assymptotically as $n \to \infty$ with $q = (n_r - 1)(n_c - 1)$ degrees of freedom.

2. Decide on level of significance: α

3. Compute p-value:

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

4. Reject Null Hypotheis:

if p-value is less than α