MATH 205: Statistical methods

Lab 9: Hypothesis testing

Hypothesis testing

In a hypothesis-testing problem, there are two contradictory hypotheses under consideration

- The null hypothesis, denoted by H_0, is the claim that is initially assumed to be true
- The alternative hypothesis, denoted by H_a, is the assertion that is contradictory to H_0.
- The null hypothesis will be rejected in favor of the alternative hypothesis only if sample evidence suggests that H_0 is false.
- If the sample does not strongly contradict H_0, we will continue to believe in the probability of the null hypothesis.

Test about a population mean

- Null hypothesis
$$H_0 : \mu = \mu_0$$
- The alternative hypothesis will be either:
  - $H_a : \mu > \mu_0$
  - $H_a : \mu < \mu_0$
  - $H_a : \mu \neq \mu_0$

Note: μ_0 here denotes a constant, and μ denotes the population mean (unknown)

This lab

- Q-Q plot
- Testing of the population mean: t-test
- Testing about the mean of two populations
- Testing about goodness of fit

1. Q-Q plot

- a Q–Q (quantile-quantile) plot is a probability plot for comparing two probability distributions by plotting their quantiles against each other
- If the two distributions being compared are similar, the points in the Q–Q plot will approximately lie on the line $y = x$
- If the distributions are linearly related, the points in the Q–Q plot will approximately lie on a line

2. Testing with a population mean

In the lecture, we consider two statistical settings

- Simplest setting
  - Normal distribution
  - σ is known
- Large-sample setting
  - ~~Normal distribution~~
    $\rightarrow$ use Central Limit Theorem $\rightarrow$ needs $n > 30$
  - ~~σ is known~~
    $\rightarrow$ replace σ by s $\rightarrow$ needs $n > 40$

For both settings, we rely on the z-value

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} \qquad \text{or} \qquad z = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$\rightarrow$ z-tests

z-test: normal distribution with known σ

- If the null hypothesis $\mu = \mu_0$ is true, then $X_i \sim N(\mu_0, \sigma^2)$
- A sample would be more contradictory to the null hypothesis than the current sample we have if

$$\bar{X} \leq \bar{x} \qquad \text{or} \qquad \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}} \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$$

- Thus, the p-value in this case is

$$P\left[Z \leq \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right] = \Phi\left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)$$

P-values for z-tests

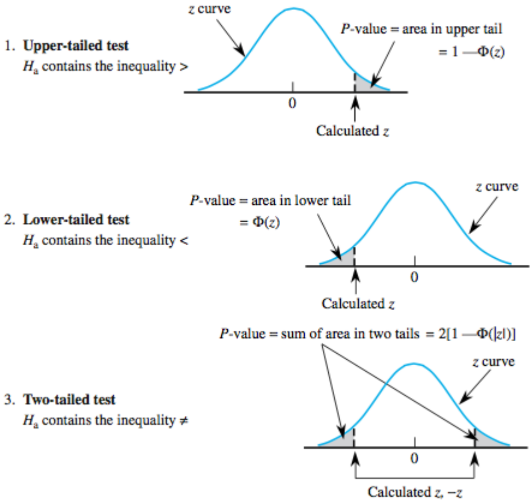


Figure 9.7 Determination of the P-value for a z test

t distributions

When $\bar{X}$ is the mean of a random sample of size n from a normal distribution with mean μ, the rv

$$\frac{\bar{X}-\mu}{S/\sqrt{n}}$$

has the t distribution with $n-1$ degree of freedom (df).

χ^2 goodness-of-fit test

- If we toss a die 150 times and find that we have the following distribution of rolls,

face	1	2	3	4	5	6
Number of rolls	22	21	22	27	22	36

  Is the die fair?

- If the die is fair, the probability of each face should be the same or $1/6$. In 150 rolls then you would expect each face to have about 25 appearances. Yet the 6 appears 36 times. Is this coincidence or perhaps something else?

χ^2 goodness-of-fit test

Idea: If we call f_i the frequency of category i, and e_i the expected count of category i, then the statistic

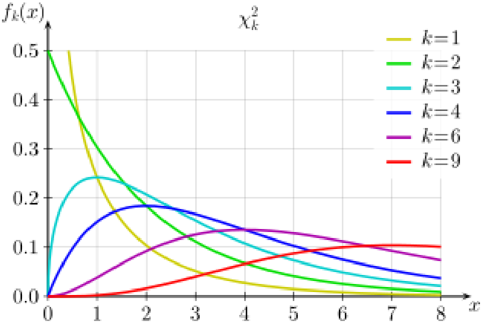
$$\sum_{i=1}^{n} \frac{(f_i - e_i)^2}{e_i}$$

follows χ^2 distribution with degree of freedom $n-1$.

Chi-squared distribution

The pdf of a Chi-squared distribution with degree of freedom ν, denoted by χ^2_ν, is

$$f(x) = \begin{cases} \dfrac{1}{2^{1/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-x/2} & x > 0 \\ 0 & x \leq 0 \end{cases}$$



Why is Chi-squared useful?

- If Z has standard normal distribution $\mathcal{Z}(0,1)$ and $X = Z^2$, then X has Chi-squared distribution with 1 degree of freedom, i.e. $X \sim \chi_1^2$ distribution.
- If $X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$ and they are independent, then

$$X_1 + X_2 \sim \chi^2_{\nu_1+\nu_2}$$

- If $Z_1, Z_2, \ldots, Z_n$ are independent and each has the standard normal distribution, then

$$Z_1^2 + Z_2^2 + \ldots + Z_n^2 \sim \chi_n^2$$

Example

- The letter distribution of the 5 most popular letters in the English language is known to be approximately

letter	E	T	N	R	O
freq.	29	21	17	17	16

  That is when either E,T,N,R,O appear, on average 29 times out of 100 it is an E and not the other 4.

- Suppose a text is analyzed and the number of E,T,N,R and O's are counted. The following distribution is found

letter	E	T	N	R	O
freq.	100	110	80	55	14

- Is this message likely to be written in English?