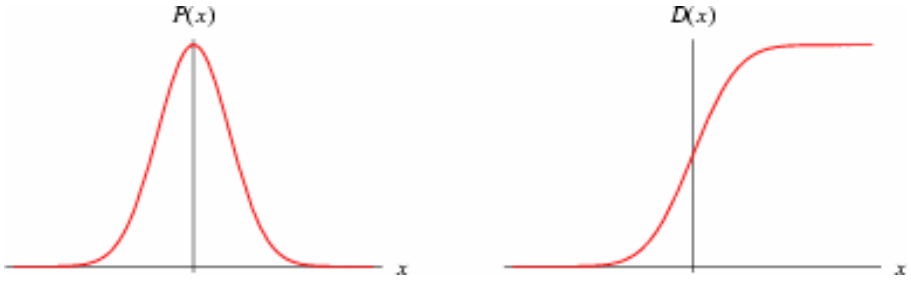
Introduction to Statistics (II)

I. Normal Distribution. A normal distribution in a variate *X* with mean μ and variance σ^2 has probability density function *P(x)* on the domain $x \in (-\infty, \infty)$. While statisticians and mathematicians uniformly use the term "normal distribution" for this distribution, physicists sometimes call it a Gaussian distribution and, because of its curved flaring shape, social scientists refer to it as the "bell curve." The probability density function (pdf) *P(x)* is

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

The so-called "standard normal distribution" is given by taking $\mu = 0$ and $\sigma = 1$ in a general normal distribution. An arbitrary normal distribution can be converted to a standard normal distribution by changing variables to $Y = \frac{X - \mu}{\sigma}$. The cumulative distribution function *D(x)*, which gives the probability that a variate will assume a value $\leq x$, is then the integral of the normal distribution,

$$D(x) = \int_{-\infty}^{x} P(t)dt = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-(t-\mu)^2/(2\sigma^2)} dt$$



II. Poisson is ultimate Binomial.

Assuming that there are *n* independent trials and success probability of each trial is *p,* the number of successful trials *X*, out of the *n* trials, then follows a binomial distribution. And

$$P(X = x) = \binom{n}{x} p^x (1-p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

If $n \to +\infty,\ p \to 0,$ and $np \to \mu$ with $\mu > 0$, then

$$\begin{aligned}\lim_{n\to+\infty} P(X=x) &= \lim_{n\to+\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\ &= \lim_{n\to+\infty} \left\{ \frac{1}{x!}(np)((n-1)p)...((n-x+1)p)(1-\frac{\mu}{n})^{n-x} \right\} \qquad (1) \\ &= \frac{\mu^x e^{-\mu}}{x!}.\end{aligned}$$

III. Test of Significance

Was it due to chance, or something else? Statisticians have invented "tests of significance" to deal with this sort of question. Nowadays, it is almost impossible to read a research article without running across tests and significance levels.

For example, assume that we observed 9 heads in 10 tosses of a coin. Since 9 is much different to what we have expected (5 heads), we would like to ask "is the coin a fair coin, or not?" There are then two statements about the coin – null hypothesis and alternative hypothesis. Each hypothesis represents one side of the argument:

- Null hypothesis—the coin tossed is a fair coin.
- Alternative hypothesis—the coin tossed is NOT a fair coin.

The null hypothesis expressed the idea that "the outcome of 9 heads in 10 tosses" is due to chance, or say the difference is due to chance. The alternative hypothesis says that "the outcome of 9 heads in 10 tosses" is due to the unfairness of the coin, or say the difference is real.

A test statistic is used to measure the difference between the data and what is expected on the null hypothesis. The *P*-value of a test is the chance of getting a big test statistic—assuming the null hypothesis to be right. *P* is not the chance of null hypothesis being right. If the *P*-value is small, say it is less than 0.05, we would reject the null hypothesis.

For the above example, we can define the test statistic as |X-5|, where X is the number of heads we observe in 10 tosses of a coin. Then the observed test statistic is 4. And the *P*-value is the probability of the event $|X-5| \geq 4$ when the coin is fair:

$$\begin{aligned}P(|X-5| \geq 4) &= P(X=0) + P(X=1) + P(X=9) + P(X=10) \\ &= (1/2)^{10} + 10(1/2)^{10} + 10(1/2)^{10} + (1/2)^{10} \\ &= 0.02\end{aligned}$$

This means that if the coin is fair, there is only about 2% probability to observe $|X-5| \geq 4$. So we would prefer to believe that the coin is NOT fair.

Z-test, t-test, and confidence interval

The choice of test statistics depends on the model and the hypothesis being considered.

One Sample Z-test Statistics

The one sample *z*-test statistics is:

$$z = \frac{observed - expected}{standard\ error}.$$

z says standard errors away an observed values is from its expected value, where the expected value is calculated using the null hypothesis. *z*-test is often used as a hypothesis testing for the mean of one sample with known variance or large sample size. The observed significance level is the chance of getting a test statistic as extreme as, or more extreme than, the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller the chance is, the stronger the evidence against the null. As *z* statistics is approximately standard normal distributed, we can use the normal distribution to compute the significance level of the *z* value we obtained.

For example:

We toss a coin 10,000 times and we observe 5,167 heads. We want to know whether the chance of heads equal to 50%? Or whether the coin is a fair coin?

Null hypothesis: the coin we tossed is a fair coin.

The observed chance of heads equal to 5,167/10,000=0.5167. The expected chance of heads for a fair coin is 0.5.

Let $x_i = 1$ for a head at the *i*th toss, $x_i = 0$ otherwise. Let $\bar{x} = \sum_{i=1}^{10,000} x_i / 10,000$. Then the standard error for the head chance estimate ($\bar{x}$) is:

$$SE = \sqrt{\frac{s^2}{n}} = \frac{\sqrt{\sum_{i=1}^{10,000} (x_i - 0.5167)^2 / (10,000 - 1)}}{\sqrt{10,000}} = 0.00499746$$

So we have $z = \frac{0.5167 - 0.5}{0.00499746} = 3.34$.

As *z* is approximately following *N*(0,1), P(*z*>3.34)=0.0004. So we reject the null hypothesis – There are too many heads to explain as chance variation.

Note that the SE can also be calculated in another way as we know the variance of observing a head under the null hypothesis:

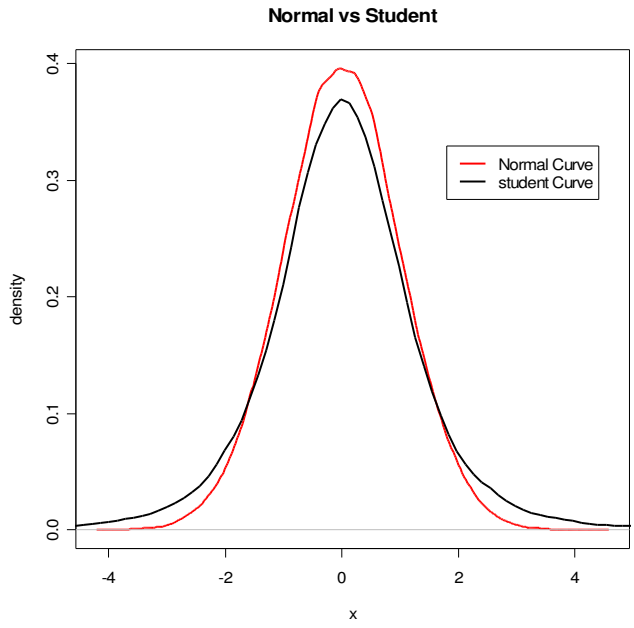
$$\sigma = \sqrt{Var(X)} = \sqrt{E(X - E(X))^2} = \sqrt{E(X - 0.5)^2} = \sqrt{0.5(1-0.5)^2 + 0.5(0-0.5)^2} = 0.5.$$

Then $SE = \sqrt{\frac{\sigma^2}{n}} = 0.005$ and $z = \frac{0.5167 - 0.5}{\sigma / \sqrt{10,000}} = \frac{0.5167 - 0.5}{0.005} = 3.34$. This tells that when sample is large, the sample variance $s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$ is a good approximation to population variance σ^2. That is why "*z*-test is often used as a hypothesis testing for the mean of one sample with known variance or large sample size."

The student's t-TEST

t-test is often used as a hypothesis testing for the mean of one sample with small sample size and without known variance. *t*-test was invented by W. S. Gossett (England, 1876-1937). Gossett worked as an executive in the Guinnes Brewery, where he went after taking his degree at Oxford. He published under the pen name "Student" because his employers did not want the competition to realize how useful the results could be.

Student's *t*-test ($z=\frac{observed\ -\ expected}{standard\ error}$) deals with the problems associated with inference based on "small" samples: the calculated mean ($\overline{X}$) and standard error may by chance deviate from the "real" mean and standard deviation (σ), or in other words, the sample variance $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2 / (n-1)$ is not a good estimate for σ^2 when the sample size *n* is small. In this situation, $z = \frac{observed - expected}{standard\ error}$ is not approximately normally distributed. Instead, it is *t*-distributed by W. S. Gossett.



In above figure, the black line is Student's curve for 4 degrees of freedom. The red line is a normal curve for comparison. The only parameter in the *t*-distribution is the degrees of freedom. In the present context, degrees of freedom = sample size -1 = $n - 1$. As *n* goes to infinity, the *t*-distribution converges to the standard normal distribution.

From above, even we have the same formula of Statistics for *z*-test and *t*-test, the significance level will be determined by different distribution curves under different circumstances. For a large sample, a normal curve is used to determine the significance level. For a small sample, a student curve is used to determine the significance level.

Confidence Intervals

A confidence interval is a range of values that has a high probability of containing the parameter being estimated. The 95% confidence interval is constructed in such a way that 95% of such intervals will contain the parameter.