Binomial Distribution

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

where value of X denotes the number of successes in n trials.

Poisson distribution

Let X be a discrete random variable taking on the values $0, 1, 2, \ldots$ such that the probability function of X is given by:

$$f(x) = P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

$$E[X] = \sum x P(x) = \sum_{k \geq 0} k \frac{1}{k!} \lambda^k e^{-\lambda}$$

$$= \lambda e^{-\lambda} \sum_{k \geq 1} \frac{1}{(k-1)!} \lambda^{k-1}$$

$$= \lambda e^{-\lambda} \sum_{j \geq 0} \frac{\lambda^j}{j!} \quad \text{by letting } j = k-1$$

But $\sum_{j\geq 0} \frac{\lambda^j}{j!} = e^{\lambda}$ (power series expansion)

$\Rightarrow E[X] = \lambda e^{-\lambda} e^{\lambda} = \lambda \Rightarrow \mu_X = \lambda$

Similarly, one can show that $\sigma_X^2 = \lambda$ so that $\sigma_X = \sqrt{\lambda}$.

If n is large ($n \geq 50$) while $np < 5$, the Binomial distribution is closely approximated by the Poisson distribution. "for rare events"

Ex) If the probability that an individual will suffer a bad reaction from injection of a flu vaccine is 0.001. Find the probability that from $n=2000$ individuals, (a) exactly 3 and (b) more than 2 individuals will suffer a bad reaction.

(a) use binomial: $P(X=3) = \binom{2000}{3}(0.001)^3(1-0.001)^{1997}$

use Poisson approximation (ok since n is large and p small):

$\lambda = np = 2000(0.001) = 2$

$P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!} \Rightarrow P(X=3) = \frac{2^3 e^{-2}}{3!} \approx 0.180$

$$P(X>2) = 1 - [P(X=0) + P(X=1) + P(X=2)] = 1 - P(X\le 2)$$

$$= 1 - \left[\frac{2^0 e^{-2}}{0!} + \frac{2^1 e^{-1}}{1!} + \frac{2^2 e^{-2}}{2!}\right]$$

$$= 1 - 5e^{-2} \approx 0.323$$

Normal Distribution

density function given by:

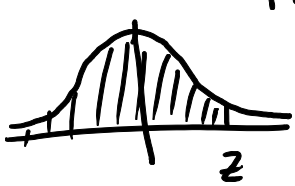
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \text{for } -\infty < x < \infty$$

$$P(X \le x) = \int_{-\infty}^{x} f(x)\,dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(v-\mu)^2}{2\sigma^2}}\,dv$$

$z = \frac{x-\mu}{\sigma}$ is std normal (mean zero, $\sigma_z = 1$)

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$ symmetric wrt to $z=0$

$$P(Z \le z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-u^2/2}\,du = \underbrace{\frac{1}{2}}_{\text{area under left half}} + \underbrace{\frac{1}{\sqrt{2\pi}} \int_{0}^{z} e^{-u^2/2}\,du}_{\text{given by table}}$$



If n is large and if neither p nor q is too close to zero, the binomial distribution can be closely approximated by normal distribution with std normal r.v.:

$$Z = \frac{X - np}{\sqrt{npq}}$$

For X binomial r.v.: $np > 5$, $nq > 5$

$$E[X] = m_x = np$$

$$\sigma_x = \sqrt{npq}$$

Ex) A fair coin is tossed 500 times. Find the probability that the number of heads will not differ from 250 by more than 10.

$$P(240 < \underbrace{X_b}_{\text{binomial}} < 260) \approx P(239.5 < \underbrace{X_n}_{\substack{\text{normal with} \\ \text{continuity correction}}} < 260.5)$$

$$= P\left(\frac{239.5 - 250}{11.18} < Z_n < \frac{260.5 - 250}{11.8}\right)$$

where $m = np = (500)\left(\frac{1}{2}\right) = 250$ and

$$\sigma = \sqrt{npq} = \sqrt{500\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)} \simeq 11.18$$

$$= P(-0.94 < Z < 0.94) = 2P(0 < Z < 0.94)$$

$$= 2A(0.94) = 2(0.3264) = 0.6528$$

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be independent random variables from some probability distribution with mean μ and variance σ^2. Then if

$S_n = X_1 + X_2 + \ldots + X_n$ then: $E[S_n] = E[X_1] + \ldots + E[X_n] = \mu + \ldots + \mu = n\mu$

$$\lim_{n \to \infty} P\left(a \leq \frac{S_n - n\mu}{\sigma\sqrt{n}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-u^2/2}\, du$$

For sample means:

if n large ($n > 30$) then distribution of sample means $\{\bar{x}\}$ for <u>all</u> possible samples of size n is approximately normal with mean $\mu_{\bar{x}} = \mu$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$.

what this says: $E[\bar{x}] = \mu$ and $\sigma_{\bar{x}}$ decreases with sample size. So values of sample mean cluster more and more closely around population mean as n increases.

confidence interval the probability $(1-\alpha)$ that the interval actually does contain the population parameter, assuming that the estimation process is repeated a large number of times. $\alpha \in (0,1)$

An interval estimate of θ is an interval of the form $\hat{\theta}_1 < \theta < \hat{\theta}_2$ where $\hat{\theta}_1$ and $\hat{\theta}_2$ are appropriate values of r.v. θ s.t:

$$P(\hat{\theta}_1 < \theta < \hat{\theta}_2) = 1-\alpha \quad \text{for } \alpha \in (0,1).$$

Note: like point estimates, interval estimates of a given parameter are not unique.

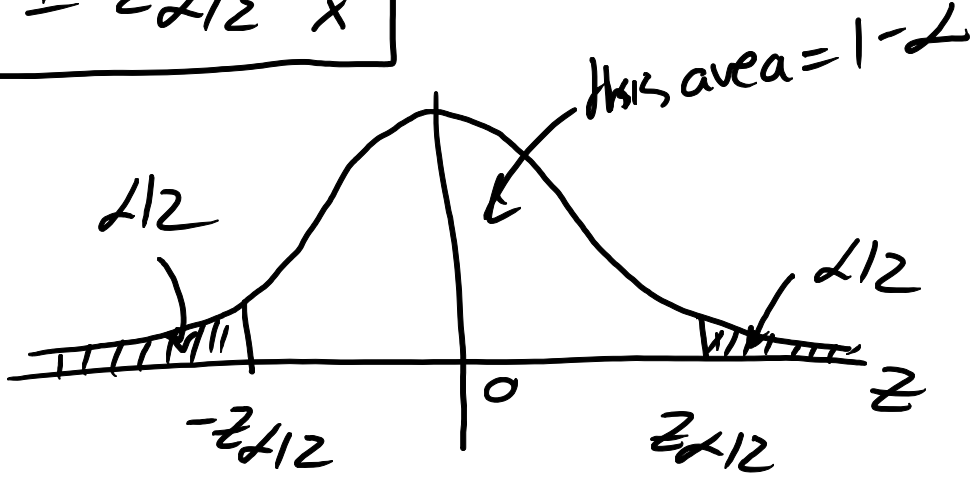
The $z_{\alpha/2}$ value

place area $\alpha/2$ in each tail.

If we place area $\alpha/2$ in each tail and if $z_{\alpha/2}$ is the value of z such that area $\alpha/2$ will lie to its right, then the confidence

interval with confidence coefficient $(1-\alpha)$ (the probability that an interval estimator encloses a population parameter) is

$$\bar{x} \pm z_{\alpha/2}\sigma_{\bar{x}}$$



$$P(|z| < z_{\alpha/2}) = 1-\alpha = P(-z_{\alpha/2} < z < z_{\alpha/2})$$

where $z = \dfrac{\bar{x}-\mu}{\sigma/\sqrt{n}} = \dfrac{\bar{x}-\mu_{\bar{x}}}{\sigma_{\bar{x}}}$

$|z| = \dfrac{|\bar{x}-\mu|}{\sigma/\sqrt{n}}$

$$\Rightarrow P\left(|\bar{x}-\mu| < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

$\Rightarrow$ <u>Thm</u> If $\bar{x}$, the mean of a random sample of size n from a <u>normal</u> population with known variance σ^2 is to be used as an <u>estimator</u> of the mean of the population,

the probability is $(1-\alpha)$ that the error will be less than $z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$.

Ex) A team of factory contractors uses the mean of a random sample of size $n=150$ to estimate the population mean of a factory product. If based on experience $\sigma=6.2$ is known (population std. deviation), what can be asserted with 0.99 probability about the max error of their estimate? Want to bound $|\bar{x}-u|$ with 99% probability.

$\Rightarrow n=150,\ \sigma=6.2,\ \alpha=0.1 \Rightarrow \frac{\alpha}{2}=0.05$

$$P(|z|<z_{\alpha/2}) = 1-\alpha = 0.99 \quad \Big| \quad z=\frac{\bar{x}-u}{\sigma/\sqrt{n}}$$

$1-0.01=0.99$

$$\Rightarrow P(-z_{\alpha/2}<z<z_{\alpha/2}) = 0.99$$

$$\Rightarrow 2P(0<z<z_{\alpha/2}) = 0.99$$

$$\Rightarrow P(0<z<z_{\alpha/2}) = \frac{0.99}{2} = 0.495 \simeq 0.4949$$

$\Rightarrow z_{\alpha/2} = 2.57$ from table

Thus, we plug into: $P(|Z| < z_{\alpha/2}) = 1-\alpha$

$$P\left(|\bar{x}-\mu| < z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha = 0.99 \quad \Rightarrow P\left(\left|\frac{\bar{x}-\mu}{\sigma/\sqrt{n}}\right| < z_{\alpha/2}\right) = 1-\alpha$$

this is the error term

$$\Rightarrow z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} = 2.57\,\frac{6.2}{\sqrt{150}} \simeq 1.30$$

Thus the team can assert with probability 0.99 that the estimation error is less than 1.30.

Note: $P(|\bar{x}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1-\alpha$

can be written as:

$$P\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$$

Thm| If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, then:

$$\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

is a $(1-\alpha)\cdot 100\%$ confidence interval for the mean of the population.

Ex) If a random sample of size $n=20$ from a normal population with variance $\sigma^2=225$ has the mean $\bar{x}=64.3$, construct a 95% confidence interval for the population mean μ. $= P(-z_{\alpha/2} < z < z_{\alpha/2})$

$\Rightarrow n=20,\ \bar{x}=64.3,\ \sigma=15$; $P(|z| < z_{\alpha/2}) = 1-\alpha$

95% interval $\Rightarrow \alpha = 0.05 \Rightarrow \alpha/2 = 0.025$

$2P(0 < z < z_{\alpha/2}) = 0.95 \Rightarrow P(0 < z < z_{\alpha/2}) = 0.475$

$\Rightarrow z_{\alpha/2} = 1.96$ (from table)

Thus, we get the interval:

$$64.3 - 1.96\frac{15}{\sqrt{20}} < \mu < 64.3 + 1.96\frac{15}{\sqrt{20}}$$

$$\Rightarrow 57.7 < \mu < 70.9$$

Ex) Use the following data (random sample) to construct a 95% confidence interval for the mean of the population sampled (assuming its normal since n small).

$$\{10, 12, 9, 6, 4, 3, 2\} \qquad n=7$$

$$\bar{x} = \frac{\Sigma x}{n} = 6.57 \ ; \quad s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}} \approx 3.9$$

$z_{.025} = 1.96$ (from table, as before)

$$\Rightarrow \quad 6.57 - 1.96\frac{3.9}{\sqrt{7}} < \mu < 6.57 + 1.96\frac{3.9}{\sqrt{7}}$$

notice that this interval would be quite large since number of samples (n) is small.

Estimation of differences between means

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

← for indep random samples from a normal population

has the standard normal distribution

$$P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha$$

Sample Sum statistics

$$Y = X_1 + X_2 + \ldots + X_n$$

X_i, indep. identically distributed from population with mean μ and variance σ^2

$$E[Y] = E[X_1] + \ldots + E[X_n] = \mu + \ldots + \mu = n\mu$$

$$Var[Y] = Var[X_1] + \ldots + Var[X_n] = \sigma^2 + \ldots + \sigma^2 = n\sigma^2$$

$$\Rightarrow std[Y] = \sigma\sqrt{n}$$