ORIGIN ≡ 0

Repeated Sampling: Distribution of Means and Confidence Intervals

Given the general setup in statistics between random variable X and the probability P(X) governed by a Probability Density Function such as the Normal Distribution, one typically uses a specific random sample to estimate the population parameters. Estimation of this sort also involves considering what happens when a population is *repeatedly sampled.* One is particularly interested in the *sampling distribution* of *repeated estimates*, such as the mean, and how these estimates may be related to probability.

For the Normal Distribution, the *population parameters* are:

μ = population mean

σ^2 = population variance

From our *sample,* we have the analogous calculations termed *point estimates*:

X_{bar} = sample mean

s^2 = sample variance

Different kinds of statistical theory underlie point estimates generally allowing them to be categorized in one of two ways:

- **"minimum variance", also known as "least squares minimum" "unbiased" or "Normal theory" estimators, and**

- **"maximum liklihood" estimators.**

How to calculate estimators of these two types is beyond the scope of introductory statistics courses. The important thing to remember is that the two methods of estimation often, but not always, yield the same point estimators. The point estimators, then feed into specific statistical techniques. Thus, it is sometimes important to know which estimator is associated with a particular technique so as not mix approaches. Maximum liklihood estimators, based on newer theory, are often specifically indicated as such (often using 'hat' notation).

In the case of estimating parameters for the Normal Distribution, X_{bar} is the point estimate for μ under both estimation theories. However s^2 sum of squares with (n-1) as divisor is the point estimate using Normal theory whereas σ^2_{hat} with same sum of sqares but using (n) as divisor is the point estimate using "maximum liklihood" theory. Confusing, yes, but now that you know the difference not all that bad...

Estimating error on point estimates of the mean:

Although X_{bar} is our Normal theory estimate of population parameter μ based on a single sample, one might readily expect X_{bar} to differ from sample to sample, and it does. Thus, we need to estimate how much X_{bar} will vary from sample to sample. Multiple sampled *means* differ from each other much less than individual sample values of X will. The relationship is called the *standard variance of the mean.* The square root of variance for the mean is called the *standard error of the mean* or simply *standard error*.

Standard Variance of the Mean = sample variance/n

or

***Standard Error of the Mean* (SEM) = sample standard deviation / $\sqrt{n}$**

Central Limit Theorem:

This result is one of the reasons why Normal theory, and the Normal Distribution underlie much of "parametric" statistics. It says that although the populations from which random variable X are drawn *may not necessarily be* normally distributed, the population of means derived by replicate sampling *will be* normally distributed. This result allows us to use the Normal Distribution with parameters μ, σ^2 estimated respectively by X_{bar} and s^2 (or occasionally σ^2_{hat}) to estimate probabilities of *means* P(X) for various values of X.

Statistics evaluating location of the mean:

Suppose we collect a sample from a population and calculate the mean X_{bar}. How reliable is X_{bar} as an estimate of μ? The usual approach is to estimate a difference (also called a distance) between X_{bar} and μ scaled to the variability in X_{bar} encountered from one sample to the next:

$$Z := \frac{X_{bar} - \mu}{\frac{\sigma}{\sqrt{n}}}$$ **< distance divided by Standard Error of the Mean**

If somehow we know the population parameter σ then we can resort directly to the standardized Normal Distribution ~N(0,1) to calculate probabilities P(Z) or cumulative probabilities Φ(Z) . However, in real life situations, σ is not known and we must estimate σ by s. When we do this, the analogous variable t:

$$t := \frac{X_{bar} - \mu}{\frac{s}{\sqrt{n}}}$$ **< Same standardizing approach but using s instead of σ**

is no longer Normally distributed. Instead, we resort to a new probability density function, known as "Student's t" to calculate P(t) or Φ(t) given t. Student's t is a commonly employed statistical function ranking high in importance along with the chi-square distribution (χ^2) and the F distribution. The Student's t distribution looks very much like the Normal distribution in shape, but is *leptokurtic*. Typically in statistical software, both distributions are utilized with analogous functions. See *Biostatistics* Worksheet 070 and the Prototype in R below for them. Although Zar in Chapter 6 perfers only to talk about the Normal distribution by assuming he/we know σ, I think it may be clearer to talk about both together here. The arguments are identical with the difference between them related to whether we know σ or whether we estimate σ by s.

Prototype in R:

```
#ANALOGOUS FUNCTIONS FOR
#NORMAL AND T DISTRIBUTIONS

#NORMAL DISTRIBUTION
mu=0       #parameter for mean
sigma=1    #paramater for standard deviation
n=1000     #number of randomly generated data points
X=1.96     #quantile X
P=0.95     # cumulative probability phi(X)
rnorm(n,mu,sigma)   #to generate random data points
dnorm(X,mu,sigma)   #P(X) from X
pnorm(X,mu,sigma)   #phi(X) from X
qnorm(P,mu,sigma)   #X from phi(X)
```

```
#t DISTRIBUTION
df=5       #degrees of freedom parameter
n=1000     #number of randomly
generated data points
X=1.96     #quantile X
P=0.95     # cumulative probability phi(X)
rt(n,df)   #to generate random data points
dt(X,df)   #P(X) from X
pt(X,df)   #phi(X) from X
qt(P,df)   #X from phi(X)
```

Confidence Interval for the Mean:

A sample Confidence Interval (CI) for a sample mean of X (or equivalently in Z or t) is the estimated range over which repeated samples of X_{bar} (or Z_{bar} or t_{bar}) are expected to fall (1-α)x100 % of the time. If a hypothesized value for mean, say μ_0, falls within a CI, then we say μ_0 is "enclosed" or "captured" by the CI with a confidence of (1-α). Equivalently, for repeated samples, μ_0 will be enclosed within repeated CI's (1- α)x100 percent of the time.

Let's calculate CI from a pseudo-random example:

$X := rnorm(100, 50, \sqrt{100})$ **< here in fact *we know* μ=50 and $\sigma^2 = 100$**

$n := length(X)$ $n = 100$

$\mu := 50$ $\sigma := \sqrt{100}$ $\sigma = 10$ **< known population standard deviation σ**

$X_{bar} := mean(X)$ $X_{bar} = 48.4955$ **< we can also *pretend* that we *don't know* the population parameters and must use sample mean and variance instead as one usually would with real data.**

$s := \sqrt{Var(X)}$ $s^2 = 96.4487$

Calculation of Confidence Intervals:

$\alpha := 0.05$ **< We choose a limit probability allowing sample means to differ from μ α X 100 percent of the time...**

$1 - \alpha = 0.95$

^ since both the Normal Distribution and the t probability distributions are symmetrical, there are equal-sized tails above and below hypothesized or known μ. Each tail therefore has α/2 probability. This is commonly known as the *Two-Tail* case...

If μ and σ are known - the Normal Distribution Case:

$\mu = 50$ $\sigma = 10$ $n = 100$

$L := qnorm\left(\frac{\alpha}{2}, 0, 1\right)$ $L = -1.96$ $\frac{\alpha}{2} = 0.025$ **< lower limit of N(0,1) for α/2**

$U := qnorm\left(1 - \frac{\alpha}{2}, 0, 1\right)$ $U = 1.96$ $1 - \frac{\alpha}{2} = 0.975$ **< upper limit of N(0,1) for α/2**

$CI := \left(\mu + L \cdot \frac{\sigma}{\sqrt{n}} \quad \mu + U \cdot \frac{\sigma}{\sqrt{n}}\right)$ **< calculating Confidence Interval using population μ and σ. Note here that I calculated each tail explicitly so I *added* both L and U to determine the CI. However, since the distribution is symmetrical, one might alternatively use:**

$CI = (48.04 \quad 51.96)$

C = the absolute value of L or U.

In that case one *subtracts* $C \cdot \frac{\sigma}{\sqrt{n}}$ from the mean for the lower limit and *adds* $C \cdot \frac{\sigma}{\sqrt{n}}$ to the mean for the upper limit.

Note here that Error of the Mean is derived from known population parameters.

If μ and σ are unknown - the t Distribution Case:

Parameters μ and σ must be estimated by sample X_{bar} and s:

$X_{bar} = 48.4955$ $s = 9.8208$

$df := n - 1$ $df = 99$

< single parameter of Student's t distribution called "degrees of freedom" df = (n-1) where n is sample size.

$L := qt\left(\frac{\alpha}{2}, df\right)$ $L = -1.9842$ $\frac{\alpha}{2} = 0.025$

$U := qt\left(1 - \frac{\alpha}{2}, df\right)$ $U = 1.9842$ $1 - \frac{\alpha}{2} = 0.975$

$CI := \left(X_{bar} + L \cdot \frac{s}{\sqrt{n}} \quad X_{bar} + U \cdot \frac{s}{\sqrt{n}}\right)$

$CI = (46.5468 \quad 50.4441)$

< calculating Confidence Interval. Note here that I calculated each tail explicitly so I *added* both L and U to determine the CI. Note also SEM is measured by the sample quantity $\frac{s}{\sqrt{n}}$

Prototype in R:

```
#CONFIDENCE INTERVALS
mu=50
sigma=10
n=100

X=rnorm(100,mu,sigma)
alpha=0.05

#NORMAL DISTRIBUTION
L=qnorm((alpha/2),0,1)
L
U=qnorm((1-alpha/2),0,1)
U
#confidence interval:
mu+L*(sigma/sqrt(n))
mu+U*(sigma/sqrt(n))

#t DISTRIBUTION
df=n-1
s=sqrt(var(X))
L=qt((alpha/2),df)
L
U=qt((1-alpha/2),df)
U
#confidence interval:
mu+L*(s/sqrt(n))
mu+U*(s/sqrt(n))
#NOTE: These values don't match MathCad
#because they are based on a different sample!
```