Central Limit Theorem and Confidence Intervals

Mark Huiskes, LIACS
mark.huiskes@liacs.nl

Universiteit Leiden

Introduction

- [Last time we have seen that the sample mean converges to the true mean for sufficiently large samples.
- Today we consider the Central Limit Theorem which tells us still a bit more: namely that the sample mean becomes normally distributed for sufficiently large samples
- Today we will not focus so much on the proof of the theorem, but rather on what we can do with it:]
- Applications of the Central Limit Theorem:
  - Approximate distributions of sums of random variables, in particular the binomial distribution
  - Construct a confidence interval for the sample mean

Central Limit Theorem for Discrete Independent Trials

- n independent trials: X1, .., Xn; E(Xi)=mu, V(Xi) = sig^2.
- [First we look at sums, later at the sample mean.] Consider the sum S_n = X_1 + … + X_n
- [Expectation=mean: sum of the expected values]

  E(S) = E(X_1) + … + E(X_n) = n mu
- Variance (because of independence of the X's):

  V(S) = V(X_1) + … + V(X_n) = n sigma^2
- Central limit theorem: Sn has, approximately, a normal density.
- "Problem 1": every S_n will have a different mean and variance: which both get large(r and larger)
- [Not a big problem, but] Solution: use standardized sums:

  S^*_n = (S_n – n mu) / sqrt(n sigma^2)

  S^*_n has E(S^*_n)= 0 and D(S^*_n) = 1 for all n (SHOW; and it will approach a standard normal density)
- If S_n = j then S^*n = x_j = (j – n mu) / sqrt(n sigma^2)

Going from discrete to continuous

- “Problem 2”: S^*_j is discrete (possible values x_j); normal density is continuous.
- Draw a figure: divide continuous axis into discrete bins. Indicate distance apart. Refer to figure 9.2 and 9.3
- Area under the histogram: eps = 1 / sqrt(n sig^2) sum_k b(n, p, k) = 1 / sqrt(n sig^2) (=distance between two spikes!)
- So solution: multiply the heights of the spikes by 1/eps
- CLT:

  P(S_n = j) \approx phi(x_j) / sqrt(n sig^2)

  where x_j = (j – n mu)/sqrt(n sig^2) and phi(x) is the standard normal density 1/sqrt(2pi) e^(-1/2 x^2)

Probability for an interval

- P(i <= S_n <=j) = P((i – mu)/sig sqrt(n) <= S^*_n <= (j – mu)/…)
- So we take: \int_i*^j* phi(x) dx
- Note from the image we can see it’s better to take (i-1/2) to (j+1/2). This is called a continuity correction.

Example

- Throw a die 420 times. S_420 = X_1 + … X_420
- What is P(1400 <= S_420 <= 1550)?
- E(X) = 3.5; V(X) = 35/12
- E(S_420) = 420 * 3.5 = 1470; V(S_420) = 420 * 35 / 12 = 1225; sig(S_420) = 35.
- P(1400<= S_420 <=1500) ~ P((1399.5 -1470) / 35 <= S*_420 <= (1550.5 -1470) / 35) = P(-2.01 <= S*_420 <= 2.3) ~NA(-2.01, 2.30)=.9670.

Approximating the Binomial Distribution

- Example: Bernoulli Trials S_n = X_1 + … + X_n.
- X=1 for succes, with probability p, X=0 for failure (prob q = 1-p)
- S_n has a binomial distribution b(n,p,k) with mean np and variance npq.
- 1. Approximation of a single probability value:

  P(S_n = j) \approx phi(x_j) / sqrt(npq)

  phi(x) = 1\sqrt(2 pi) e^(-1/2 x^2)

- 2. Approximation of an interval:

  P(i <= Sn <= j) = \int_i*^j* phi(x) dx

  With i* = i-1/2-np/sqrt(npq) and j*=

When to use which approximation?

- Small n: just use the binomial distribution itself
- Large n, small p: use the Poisson approximation
- Large n, moderate p: use the normal density, esp accurate for values of k not too far from np.

Distribution of the Sample Mean

- [So far we have looked at sums of independent random variables. Now we will look at the sample mean. For large n also the sample mean is normally distributed]
- A_n = 1/n (X_1 + … + X_n)
- Again E(Xi) = mu, V(Xi) = sig^2. We use A_n to estimate mu
- E(A_n) = mu, V(A_n) = sigma^2 / n, D(A_n) = sigma/sqrt(n) (standard error = standard deviation of the sample mean).

- Central Limit Theorem: A_n = S_n / n has a normal density, and A^*n = (A_n – mu) / (sig/sqrt(n)) has a standard normal density.
- Show what this means. Move to paper

Confidence intervals

- Show with a picture what that means: use worked out text on paper.
- Work out the probability of P(mu – r <= A_n <= mu + r)
- A_n has a normal distribution with mean mu and standard deviation the standard error. So we can compute this probability by transforming to the standard normal density.

- Form of a confidence interval:

  best estimate +/- “some number” x standard error of best estimate

Computing confidence interval for the mean with known standard deviation

- Compute the sample mean and standard error
- Compute the z-value corresponding to the confidence level
- Confidence interval: sample mean +/- z_c * standard error.

Example

- Sample of 100 observations. Sample mean: A_n = 10. Suppose standard deviation of a measurement is known to be 2. Construct a 95% confidence interval for the sample mean.

- 95% confidence: z = 1.96.
- Confidence interval: sample mean +/- z * standard error.
- Standard error: 2 / sqrt(100) = 0.2

Confidence interval: [10 – 1.96 * 0.2, 10 + 1.96 * 0.2] = [9.61,10.39]

Unknown standard deviation

- What if we don’t know the standard deviation:
  - We simply take the sample standard error: works well if n is sufficiently large
  - For n not large, we need to use the t-distribution instead of the normal distribution