NRES 798 — Lab 6

Confidence intervals, etc.

Essentially these same calculations will be constantly used in hypothesis testing.

Confidence intervals for the mean

Generate again our sample of spider tibial spine lengths: `tibia <- rnorm(50, 0.253, 0.0039)`. Compute the (sample) mean, standard deviation, and standard error ($s/\sqrt{n}$).

Large samples

We know that the *confidence limits* (ends of the confidence interval) are of the form $\overline{Y} \pm k\,\mathrm{SE}$, for some k. Ignoring for now the variability of the estimated SE, and assuming that $\overline{Y}$ is normal, the standardized RV $Z = (\overline{Y} - \mu)/\mathrm{SE} \sim N(0, 1)$ (standard normal). Therefore, k must be such that the probability of Z being between $-k$ and k equals the desired *confidence level*, e.g., 0.95.

Sketch a standard normal PDF (or plot it with R), with $-k$ and k shown on the variable axis (Z in this case). Probability is the area under the PDF f. The CDF F is the probability (area) to the left of a point. Use the standard normal CDF to find the probability between -2 and 2.

Think about that. An interval $\overline{Y} \pm 2\,\mathrm{SE}$ will include μ with the probability that you calculated. Did your *observed* interval include the "true" μ (the μ that you used to generate the sample)?

Find the probability of μ being within $\overline{Y} \pm \text{SE}$. What about $\overline{Y} \pm 3\,\text{SE}$? It is useful to remember these probabilities for quick assessments of published SE's. At least the one for $k = 2$.

We found confidence levels for given k. To go the other way we need F^{-1}, the quantile function. Look at your sketch, how can you find k such that the probability between $-k$ and k is 0.95? Hint: use only one of k or $-k$, and add the areas to the left or right.

Use the standard normal quantil function to calculate $F^{-1}(0.025 + 0.95)$, $F^{-1}(1 - 0.025)$, $F^{-1}(0.025)$. See in your sketch if this makes sense.

Find the confidence interval for a confidence level of 0.99.

Small samples

Now, let us take into account the variability of the estimated SE. Assuming that the Y_i are normal, Z now has a t-distribution with parameter $n - 1$. Repeat the above with the t-distribution instead of the normal.

The differences are more important in smaller samples. Do the t calculations with a sample of 20 (`tibia[1:20]`, or `head(tibia, 20)`).

Calculate and/or plot the t quantile for various n.

Sample size

Roughly, 95% of the time the mean has an error of less than $\pm 2\,\text{SE} = \pm 2\sigma/\sqrt{n}$. If we have a guess for σ, what sample size do we need in order to get an error less than $\pm E$, 95% of the time? Produce a formula for n [1].

More practical: if we have a guess for the coefficient of variation, what sample size do we need in order to get an error less than $\pm E\%$, 95% of the time? Do the calculations for $\text{CV} = 30\%$ and $E = 5\%$.

[1] Standard 3-step problem-solving routine: (1) Write down what you know. (2) Figure out what you want to know. (3) Solve for it. Most of the time one should add: (0) Draw a picture.

Integration

Let's see different ways of finding the area under a curve (integral). Take the standard normal PDF between -1 and 1.

1. Divide the area into vertical strips of width Δ: Area $\approx \sum f(x_i)\Delta$. For instance, `delta <- 0.1`, `x <- seq(from=-1, to=1-delta, by=delta)`, `sum(dnorm(x) * delta)`. Do it. Try smaller and larger values of delta.

2. A more sophisticated version of the previous method: `integrate(dnorm, lower=-1, upper=1)`.

3. Monte Carlo integration. Average the height at random points and multiply by the base: `x <- runif(100, -1, 1)`, `mean(dnorm(x)) * 2`. For the same number of function evaluations, is this more or less accurate than method 1? This is more easily generalized to more variables. E.g., in two dimensions, the area under a surface, within some region of the plane. Similar methods are used a lot in Bayesian inference.

4. Since we have the CDF $F(y) = \int_{-\infty}^{y} f(x)\,\mathrm{d}x$, we can use $\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$ with $a = -1$ and $b = 1$. See your sketch. Do it.