Ch8: Confidence Intervals

- **Dataset** description due tonight 10pm
- **HW4** due Thu 10pm

4 Oct 2011
BUSI275
Dr. Sean Ho

TRINITY WESTERN UNIVERSITY

Outline for today

- Making estimates on the binomial proportion
- Confidence intervals
  - On μ, with known σ
  - On the binomial proportion π
  - On μ, with unknown σ
    - Student's *t*-distribution

Binomial sampling distribution

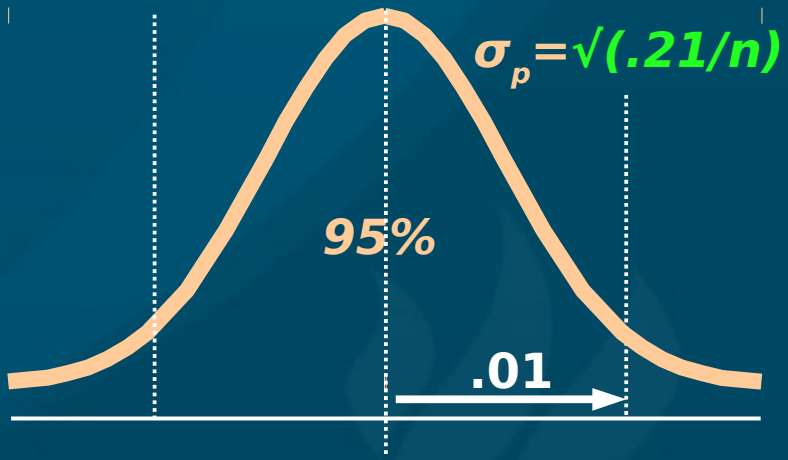
- For most (n,p), the binomial is approx. normal:
  - μ = np, σ = √(npq)
- Let π be the "true" prob of success in the pop
  - p = observed prob of success in sample
- Convert from "number of successes" (x) to "probability of success" (p):

Just divide by n (total # of trials):

	# successes	prob. of success
Mean	μ = np →	$\mu_p = \pi$
Std dev	σ = √(npq) →	$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$

Binomial example

- Assume about 70% of people like our toothpaste. We want to refine this estimate, to a precision of ±1%, with 95% confidence.
  - How many people do we need to poll?
- Prob. of success ⇒ binomial
- 95% conf. ⇒ $z = \pm 1.96$
  - NORMSINV(.025)
- Std. err $\sigma_p = \sqrt{((.70)(.30)/n)}$



- Putting it together: $1.96 = .01 / \sigma_p$.
- ⇒ $n = (1.96 / .01)^2 (.70)(.30) \approx 8068$

Confidence intervals

Tip: log(income) is often normal

- "If we were to select another random sample from the same population, 95% of the time its mean would lie between _____ and _____."
  - Application of the SDSM
- E.g., avg income of 25 students is $12,000.
  - Assume σ = $4,000 (pop. SD!)
- Std err is $\sigma_{\bar{x}} = \sigma/\sqrt{n} = \$800$
- 95% conf. ⇒ $z = \pm 1.96$
- So the confidence interval is $12k ± (1.96)(800)
  - We think the true mean income lies somewhere between $10,432 and $13,568, with 95% confidence.

Myths about confid. intervals

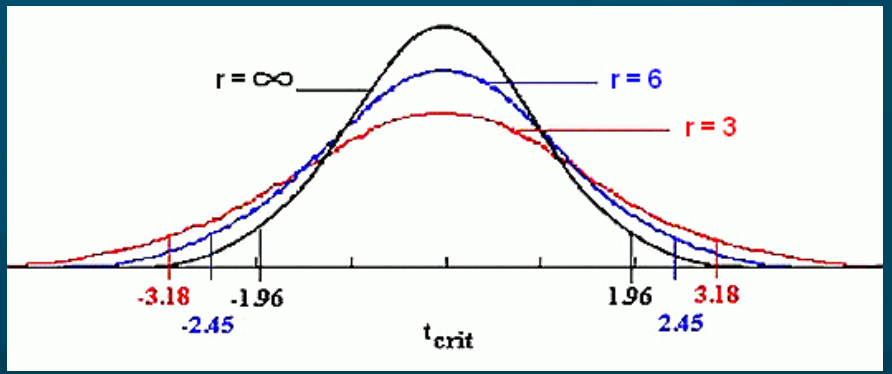
- Myth: “All students in this population have income between $10.4k and $13.5k”
- Myth: “95% of students in this population have income between $10.4k and $13.5k”
- Myth: “If we repeated the study, 95% of the students surveyed would have income betw….”
- Myth: “We are 95% sure the mean income of our sample of 25 students is between ….”

Confid. interval for binomial

- In a poll of 80 people, 60 like our product
  - Point estimate: $p = 75\%$
- Obtain a 95% confidence interval:
  - 95% confid. $\Rightarrow z = \pm 1.96$
- Std err: $\sigma_p = \sqrt{(pq/n)} = \sqrt{((.75)(.25)/80)} \approx 4.84\%$
- Put it together: (*pt estimate*) ± (*z*)(*std err*)
  - $75\% \pm (1.96)(4.84\%)$
- We are 95% confident that between 65.51% and 84.49% of people like our product
  - i.e., that the real proportion π is in that range

Confid. int., with unknown σ

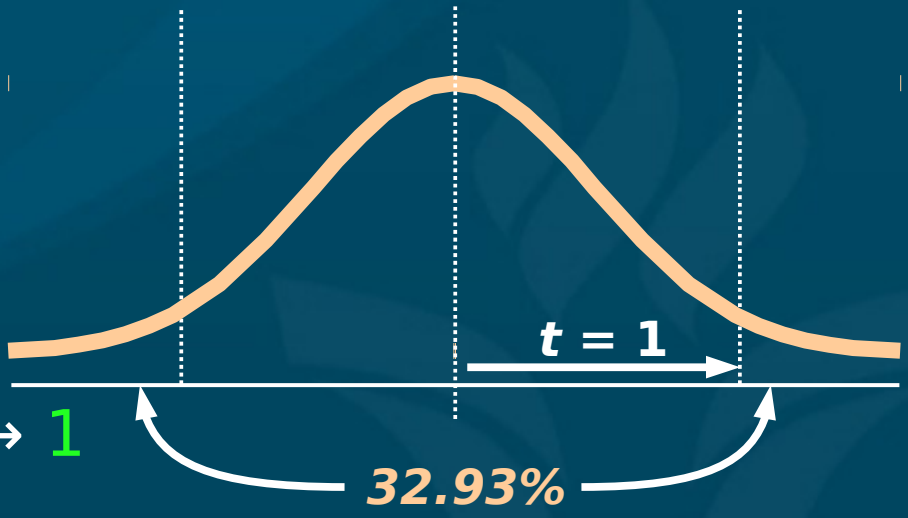
- What if we don't know the population σ?
- Estimate it from the sample SD: s
  - But this adds uncertainty in estimating μ
- Use "Student's" *t*-distribution on SDSM
  - Similar to normal, but wider (w/uncertainty)
  - Degrees of freedom: df = n-1
  - Approaches normal as df increases



William Sealy Gosset in 1908 (Wikipedia)

t-distribution in Excel

- TDIST(*t*, *df*, *tails*)
  - *t*: t-score, akin to z-score (x – μ) / SE
  - *df*: degrees of freedom, df = n-1 for now
  - *tails*: 1 for area in one tail, or 2 for both tails
  - Result: % area under the *t*-dist in tail(s)
    - TDIST(1, 20, 2) → 32.93%
- TINV(*area*, *df*)
  - Always assumes area is total in both tails
  - Result: t-score
    - TINV(0.3293, 20) → 1



Confidence interval: example

- Track sales this month at 25 stores out of 1000:
  - Average = 8000 units, SD = 1500
- Estimate the average sales this month across all 1000 stores (i.e., 95% confidence interval).
- Standard error: $s/\sqrt{n} = 1500/5 = 300$
- Only have s, not σ: so use t-dist (df=24)
  - TINV(.05, 24) → t = ±2.0639
- Putting it together: $8000 \pm (2.0639)(300)$
  - 7380.83 (round down), 8619.17 (round up)
  - With 95% confidence, the average sales this month across all stores is between 7380 and 8620 units

Project: variables & data

- Ensure your sample size is sufficient!
- Sample size = # observations
  - Not total # of numbers in the spreadsheet!
  - What is the unit of observation?
- Select fewer but more relevant variables
  - More variables = more complete model, but
  - More variables = harder for you to find significant effects during analysis
- E.g., survey with 100 questions, 20 participants:
  - Total of 2000 numbers, but sample size is only 20!

TODO

- Dataset description due tonight 10pm
- HW4 (ch5): due Thu at 10pm
  - Remember to format as a document!
  - HWs are to be individual work
- REB form due Tue 18 Oct 10pm
  - Deadline postponed a week
  - If using non-public human-subjects data, also submit printed signed copy to me
  - You may want to submit early to allow time for processing by TWU's REB (3-4 weeks)