Ch7-8: Making Inferences About the Population

- ***HW4** due Thu 10pm*
- ***Dataset** description due tonight*
- ***REB** due next Tue*

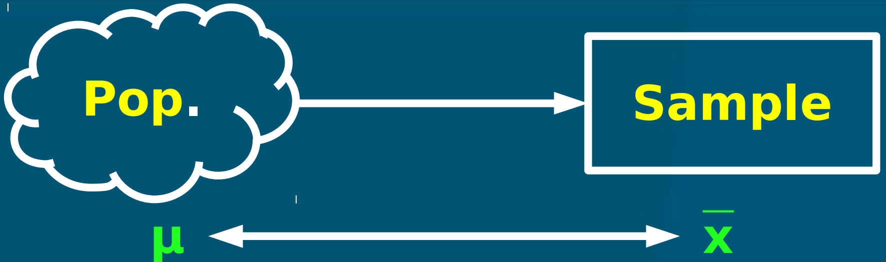
7 Feb 2012
Dr. Sean Ho

busi275.seanho.com

Outline for today

- Sampling distribution of the sample mean
  - $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$
  - Central Limit Theorem
- Uses of the SDSM
  - Probability of $\bar{x}$ above a threshold
  - Estimating needed sample size
- Estimates on a binomial proportion
- Confidence intervals
  - On μ, with known σ
  - On the binomial proportion π
  - On μ, with unknown σ (Student's t-dist)

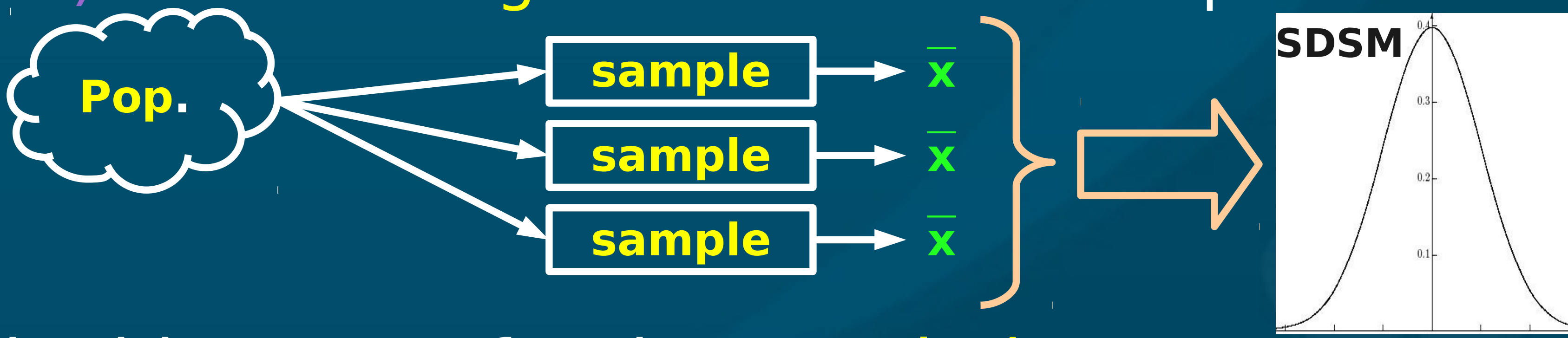
Sampling error



- Sampling is the process of drawing a sample out from a population
- Sampling error is the difference between a statistic calculated on the sample and the true value of the statistic in the population
- e.g., pop. of 100 products; avg price is $\mu=\$50$
  - Draw a sample of 10 products, calculate average price to be $\bar{x}=\$55$
  - We just so happened to draw 10 products that are more expensive than the average
  - Sampling error is \$5

Sampling distribution

1) Draw one sample of size n
2) Find its sample mean $\bar{x}$ (or other statistic)
3) Draw another sample of size n; find its mean
4) Repeat for all possible samples of size n
5) Build a histogram of all those sample means



- In the histogram for the population, each block represents one observation
- In the histogram for the sampling distribution, each block represents one whole sample!

SDSM

- Sampling distribution of sample means
  - Histogram of sample means ($\bar{x}$) of all possible samples of size n taken from the population
  - It has its own mean, $\mu_{\bar{x}}$, and SD, $\sigma_{\bar{x}}$
- SDSM is centred around the true mean μ
  - i.e., $\mu_{\bar{x}} = \mu$
- If μ=\$50 and our sample of 10 has $\bar{x}$=\$55, we just so happened to take a high sample
  - But other samples will have lower $\bar{x}$
  - On average, the $\bar{x}$ should be around \$50

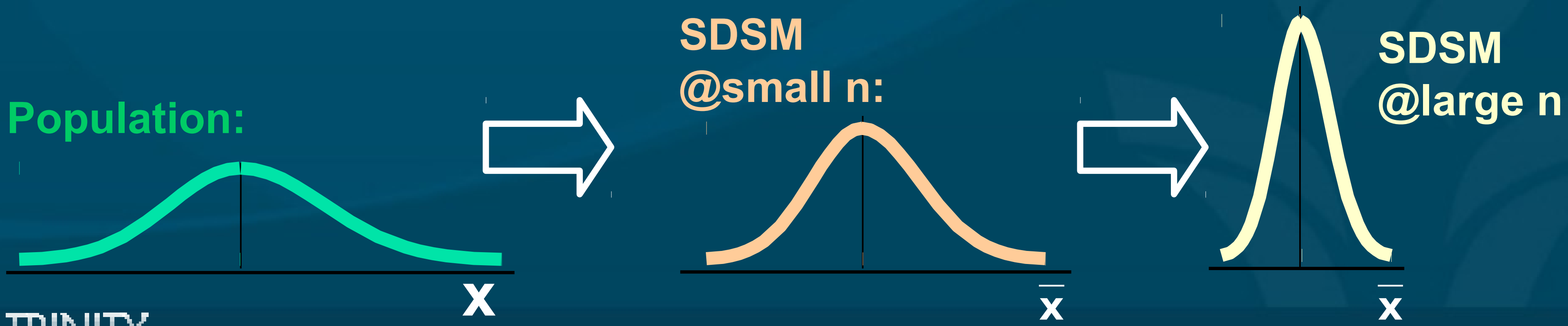
Properties of the SDSM

- $\mu_{\bar{x}} = \mu$: centred around true mean
- $\sigma_{\bar{x}} = \sigma/\sqrt{n}$: narrower as sample size increases
  - For large n, any sample looks about the same
  - Larger n ⇒ sample is better estimate of pop
  - $\sigma_{\bar{x}}$ is also called the standard error
- If pop is normal, then SDSM is also normal
- If pop size N is finite and sample size n is a sizeable fraction of it (say >5%), need to adjust standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$$

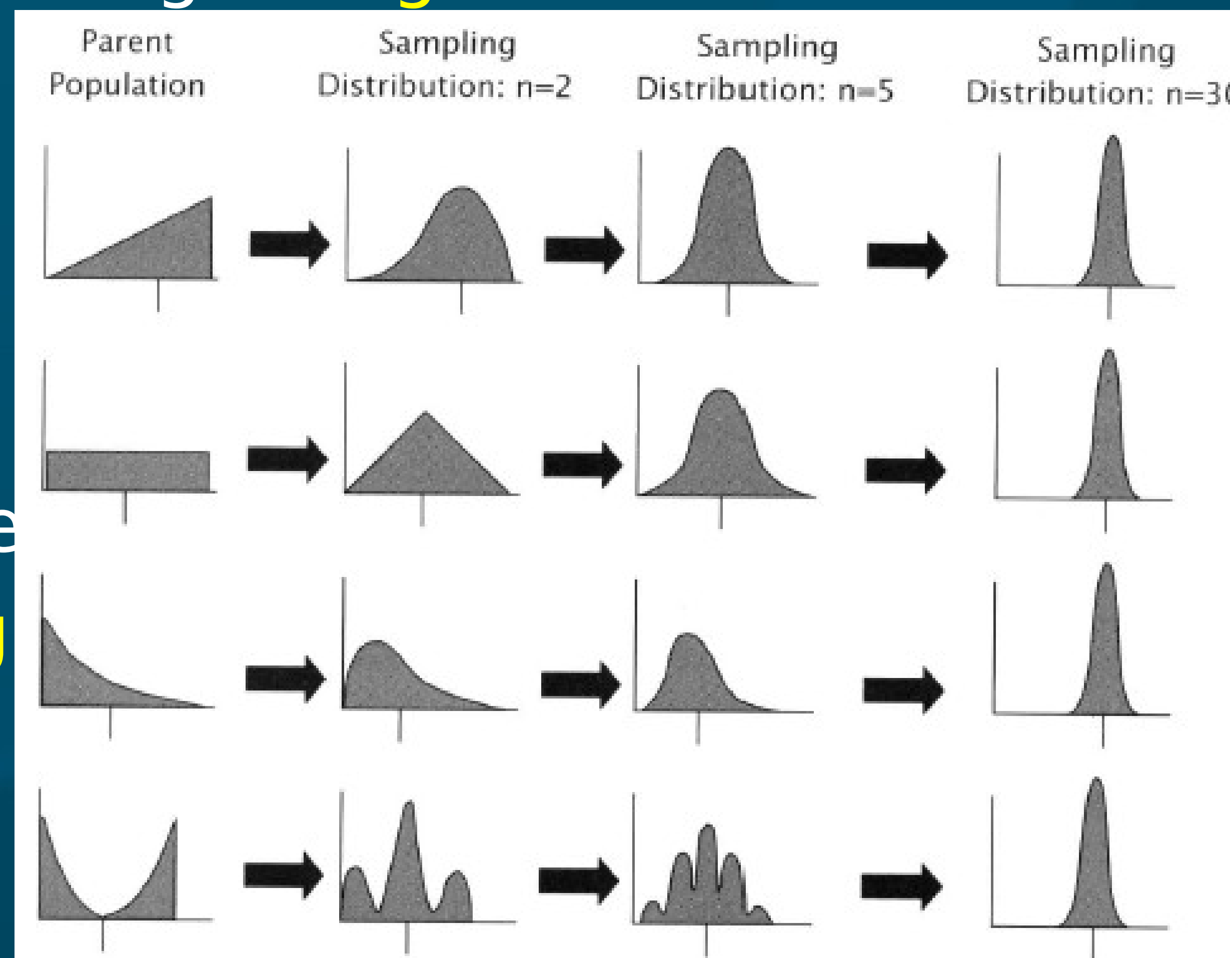
Central Limit Theorem

- In general, we won't know the shape of the population distribution, but
- As n gets larger, the SDSM gets more normal
  - So we can use NORMDIST/INV to make calculations on it
- So, as sample size increases, two good things:
  - Standard error decreases ($\sigma_{\bar{x}} = \sigma/\sqrt{n}$)
  - SDSM becomes more normal (CLT)



SDSM as n increases

- @n=1, SDSM matches original population
- As n increases, SDSM gets tighter and normal
- Regardless of shape of original population!
- Note: pop doesn't get more normal; it does not change
- Only the sampling distribution changes

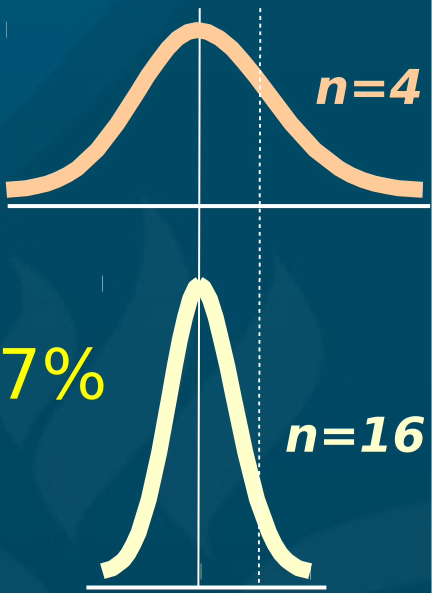


Outline for today

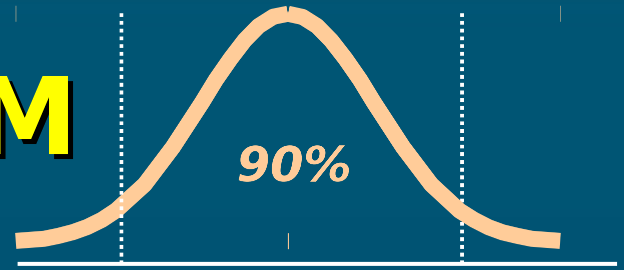
- Sampling distribution of the sample mean
  - $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$
  - Central Limit Theorem
- Uses of the SDSM
  - Probability of $\bar{x}$ above a threshold
  - Estimating needed sample size
- Estimates on a binomial proportion
- Confidence intervals
  - On μ, with known σ
  - On the binomial proportion π
  - On μ, with unknown σ (Student's t-dist)

Example 1: SDSM

- Say package weight is normal: μ=10kg, σ=4kg
  - Say we have to pay extra fee if the average package weight in a shipment is over 12kg
- If our shipment has 4 packages, what is the chance we have to pay fee?
  - Standard error: $\sigma_{\bar{x}} = 4/\sqrt{4} = 2$kg
  - $z = (\bar{x} - \mu_{\bar{x}})/\sigma_{\bar{x}} = (12-10)/2 = 1$
  - Area to right: 1-NORMSDIST(1)=15.87%
    - Or: 1 - NORMDIST(12, 10, 2, 1)
- 16 pkgs?
  - Std err: $\sigma_{\bar{x}} = 4/\sqrt{16} = 1$kg; $z = (12-10)/1 = 2$
  - Area to right: 1-NORMSDIST(2) = 2.28%



Example 2: SDSM



- Assume mutual fund MER norm: μ=4%, σ=1.8%
  - Broker randomly(!) chooses 9 funds
  - We want to say, "90% of the time, the avg MER for the portfolio of 9 funds is between ___% and ___%." (find the limits)
- Lower limit: 90% in middle ⇒ 5% in left tail
  - NORMSINV(0.05) ⇒ z = -1.645
  - Std err: $\sigma_{\bar{x}} = 1.8/\sqrt{9} = 0.6\%$
  - $z = (\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}}$ ⇒ $-1.645 = (\bar{x} - 4) / 0.6$
  - ⇒ lower limit is $\bar{x} = 4 - (1.645)(0.6) = 3.01\%$
- Upper: $\bar{x} = \mu + (z)(\sigma_{\bar{x}}) = 4 + (1.645)(0.6) = 4.99\%$

Example 2 (MER): conclusion

- We conclude that, if the broker randomly chooses 9 mutual funds from the population
- 90% of the time, the average MER in the portfolio will be between 3.01% and 4.99%
  - This does not mean 90% of the funds have MER between 3.01% and 4.99%!
  - 90% on SDSM, not 90% on orig. population
- If the portfolio had 25 funds instead of 9, the range on avg MER would be even narrower
  - But the range on MER in the population stays the same

SDSM: estimate sample size

- So: given μ, σ, n, and a threshold for $\bar{x}$
  ⇒ we can find probability (% area under SDSM)
  - Std err ⇒ z-score ⇒ % (use NORMDIST)
- Now: if given μ, σ, threshold $\bar{x}$, and % area,
  ⇒ we can find sample size n
  - Experimental design: how much data needed
- Outline:
  - From % area on SDSM, use NORMINV to get z
  - Use $(\bar{x} - \mu)$ to find standard error $\sigma_{\bar{x}}$
  - Use $\sigma_{\bar{x}}$ and σ to solve for sample size n

Example 3: min sample size

- Assume weight of packages is normally distributed, with $\sigma=1$kg
- We want to estimate average weight to within a precision of ±50g, 95% of the time
  - How many packages do we need to weigh?
- NORMSINV(0.975) → $z=\pm1.96$
  - $\pm1.96 = (\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}}$.
  - Don't know μ, but we want $(\bar{x} - \mu) = \pm50$g
  - $\Rightarrow \sigma_{\bar{x}} = 50\text{g} / 1.96$
  - So $\sigma/\sqrt{n} = 50\text{g} / 1.96$. Solving for n:
  - $n = (1000\text{g} * 1.96 / 50\text{g})^2 = 1537$ (round up)

Outline for today

- Sampling distribution of the sample mean
  - $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$
  - Central Limit Theorem
- Uses of the SDSM
  - Probability of $\bar{x}$ above a threshold
  - Estimating needed sample size
- Estimates on a binomial proportion
- Confidence intervals
  - On μ, with known σ
  - On the binomial proportion π
  - On μ, with unknown σ (Student's t-dist)

Binomial sampling distribution

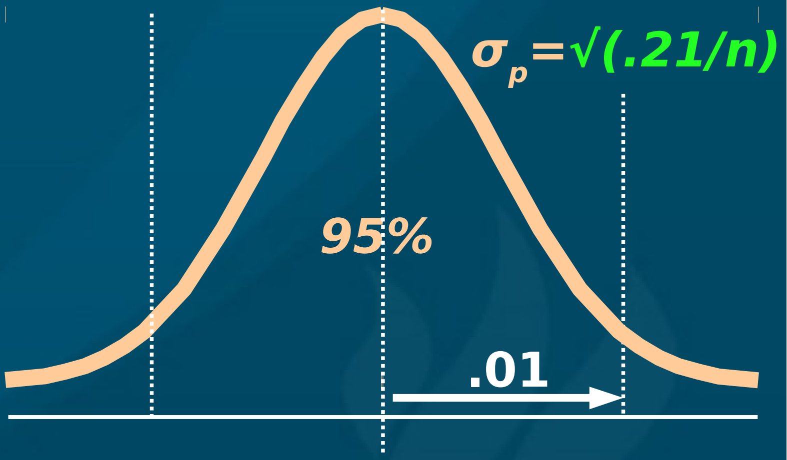
- For most (n, p), the binomial is approx. normal:
  - $\mu = np$, $\sigma = \sqrt{(npq)}$
- Let π be the "true" prob of success in the pop
  - p = observed prob of success in sample
- Convert from "number of successes" (x) to "probability of success" (p):

Just divide by n (total # of trials):

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

Example 4: Binomial, find n

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

Outline for today

- Sampling distribution of the sample mean
  - $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$
  - Central Limit Theorem
- Uses of the SDSM
  - Probability of $\bar{x}$ above a threshold
  - Estimating needed sample size
- Estimates on a binomial proportion
- Confidence intervals
  - On μ, with known σ
  - On the binomial proportion π
  - On μ, with unknown σ (Student's t-dist)

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 ….”

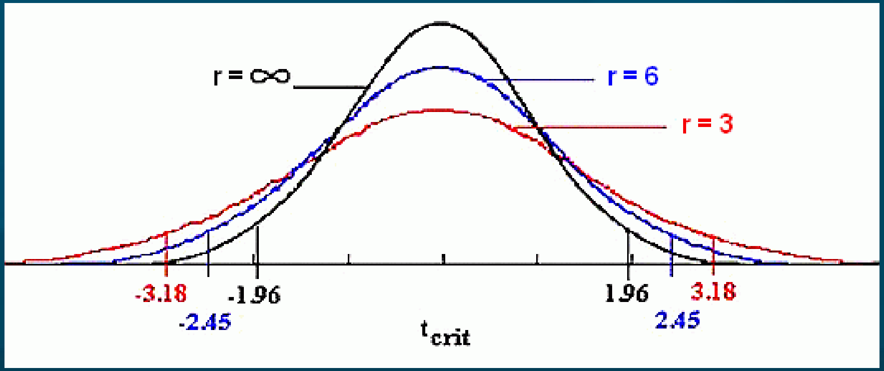
Example 5: Binomial conf int

- In a poll of 80 people, 60 like our product
  - Point estimate: $p = 75\%$
- Obtain a 95% confidence interval:
  - 95% confid. ⇒ $z = \pm 1.96$
- Std err: $\sigma_p = \sqrt{(pq/n)} = \sqrt{((.75)(.25)/80)} \approx 4.84\%$
- Put it together: (*pt estimate*) ± (*z*)(*std err*)
  - 75% ± (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

Conf int, with unknown σ

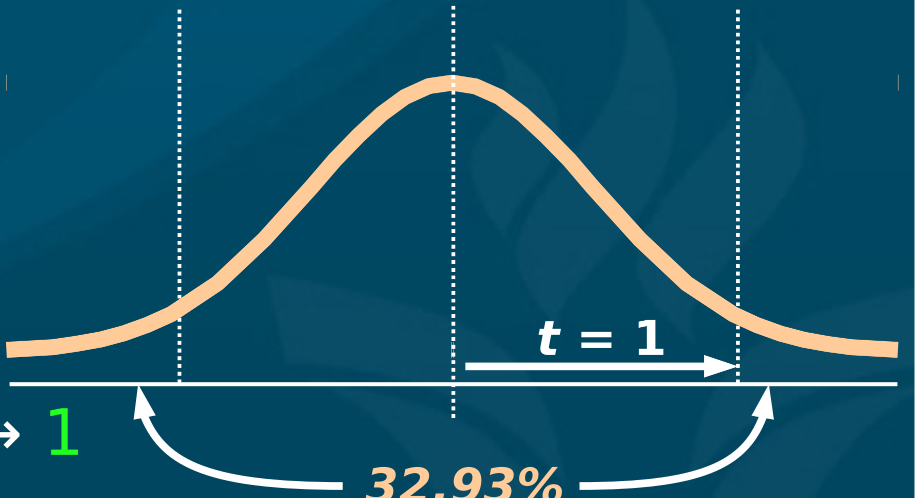
- What if we don't know the population σ?
- Estimate it from the sample SD: s
  - But this adds uncertainty in estimating μ
- Use "Student's" *t*-distribution on SDSM
  - Similar to normal, but wider (w/uncertainty)
  - Degrees of freedom: df = n-1
  - Approaches normal as df increases

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



Example 6: confidence interval

- 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 ± (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

Summary

- What is the question asking for?
  - Find percent area under SDSM (*DIST)
  - Find threshold on SDSM (*INV)
    - Find confidence interval (2 thresholds)
  - Find min required sample size (n)
- What kind of distribution?
  - SDSM, σ known (normal)
  - SDSM, σ unknown (t-dist)
  - Binomial (normal)

TODO

- HW4 (ch5-6): due this Thu 9Feb
- Dataset description tonight 10pm
  - If using existing data, need to have it!
  - If gather new data, have everything for your REB application: sampling strategy, recruiting script, full questionnaire, etc.
- REB application next week: 14Feb (or earlier)
  - If not REB exempt, need printed signed copy