Ch7: Sampling Distributions

- **HW3 due 10pm**

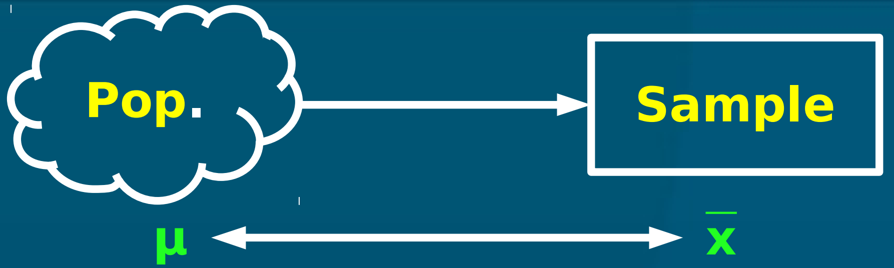
29 Sep 2011
BUSI275
Dr. Sean Ho

Trinity Western University

Outline for today

- Sampling distributions
  - Sampling distribution of the sample mean
  - $\mu_{\bar{x}}$ and $\sigma_{\bar{x}}$
  - Central Limit Theorem
- Uses of the SDSM
  - Probability of sample avg above a threshold
  - 90% confidence interval
  - Estimating needed sample size

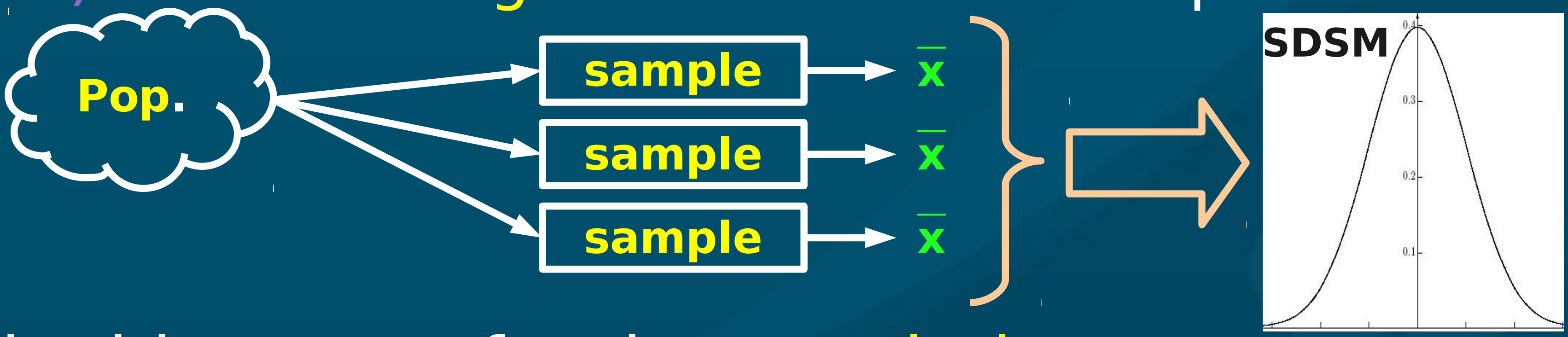
Sampling error



- 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 μ=\$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

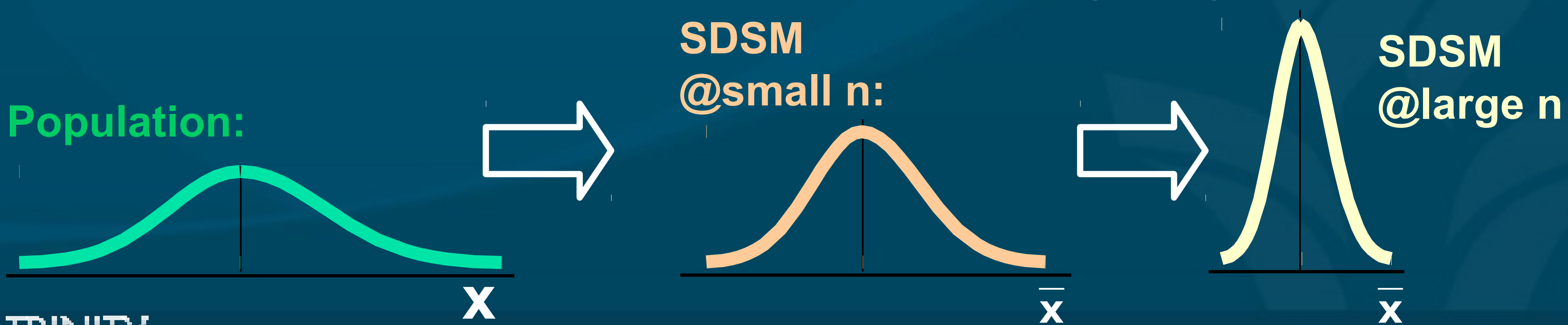
Properties of the SDSM

- $\mu_{\bar{x}} = \mu$: 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:

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

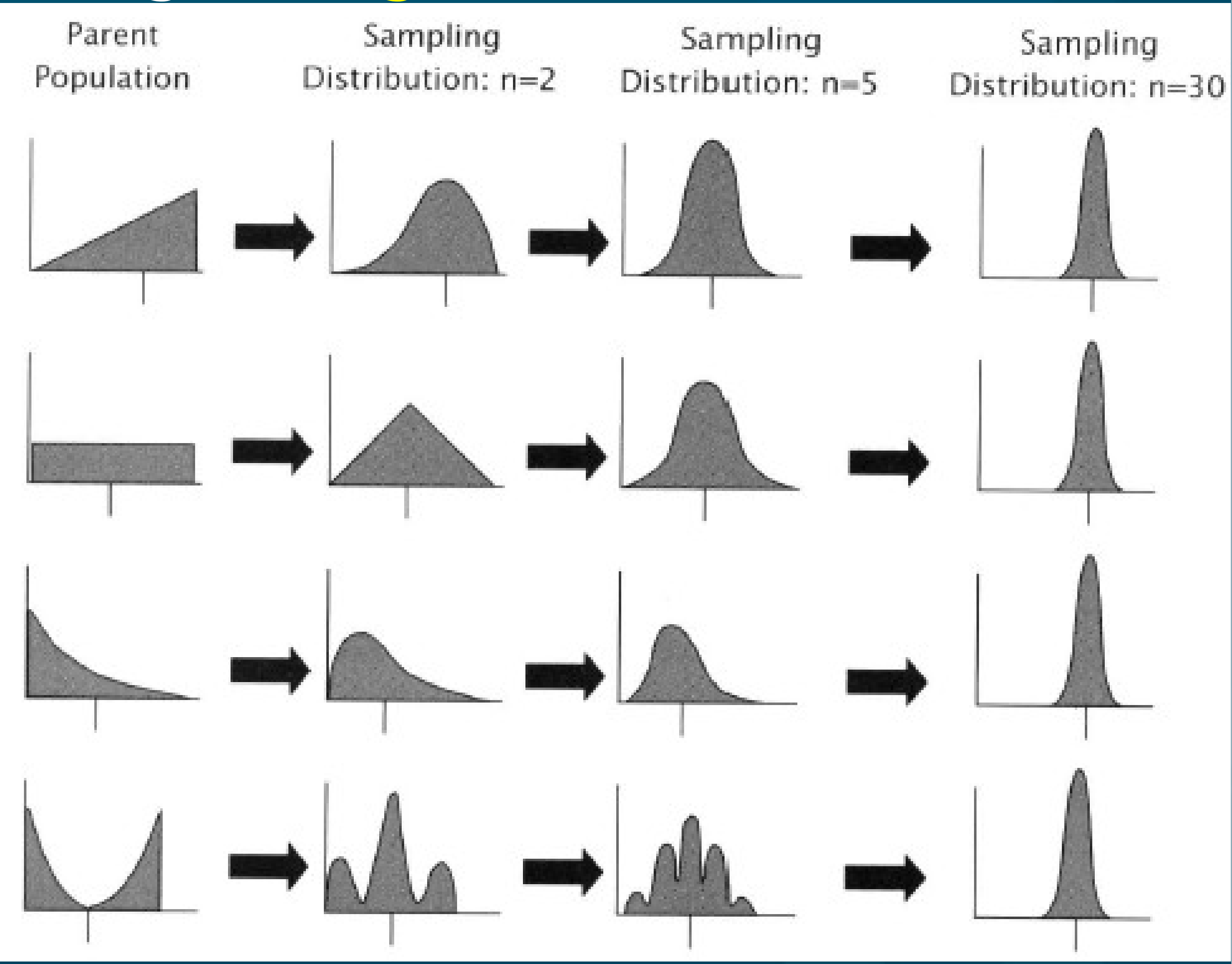
Central Limit Theorem

- 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)



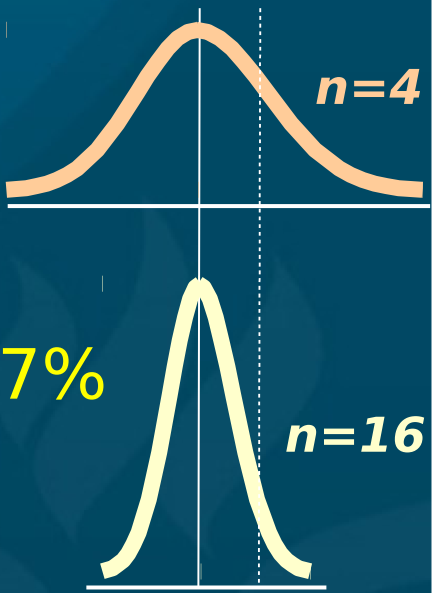
SDSM as n increases

- @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



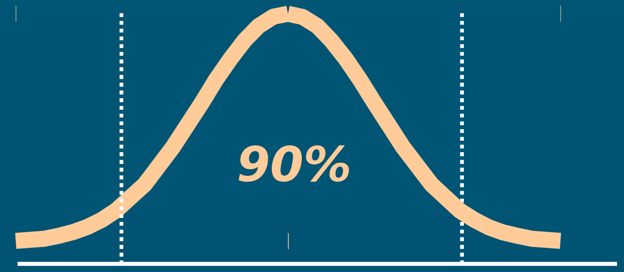
SDSM: example

- 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%

SDSM: example



- Assume mutual fund MER norm: $\mu=4\%$, $\sigma=1.8\%$
  - 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: $\sigma_{\bar{x}} = 1.8/\sqrt{9} = 0.6\%$
  - $z = (\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}} \Rightarrow -1.645 = (\bar{x} - 4) / 0.6$
  - ⇒ lower limit is $\bar{x} = 4 - (1.645)(0.6) = 3.01\%$
- Upper: $\bar{x} = \mu + (z)(\sigma_{\bar{x}}) = 4 + (1.645)(0.6) = 4.99\%$

MER example: 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

Estimating needed sample size

- Assume weight of packages is normally distributed, with $\sigma = 1\text{kg}$
- 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) → $z = \pm 1.96$
  - $\pm 1.96 = (\bar{x} - \mu_{\bar{x}}) / \sigma_{\bar{x}}$.
  - Don't know μ, but we want $(\bar{x} - \mu) = \pm 50\text{g}$
  - $\Rightarrow \sigma_{\bar{x}} = 50\text{g} / 1.96$
  - 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)

TODO

- HW3 (ch3-4): due tonight at 10pm
  - Remember to format as a document!
  - HWs are to be individual work
- Dataset description due this Tue 4 Oct