SySc 512, Statistical Inference Session 17,
[NetLogo Demo / Central Limit Theorem] (sampling)

Define: Random number (sample mean)
$\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$, w/ n = # samples

where the sequence of random variables x_i are drawn from distrib w/ mean = μ & variance = σ^2 (but distrib is unknown).

Define: Sample variance (random number)
$S_n^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$

Note: S_n^2 is a "biased estimator" of σ^2, because 2 free parameter of the data are being estimated: popul mean & variance. The "unbiased estimator" is S_{n-1}^2.

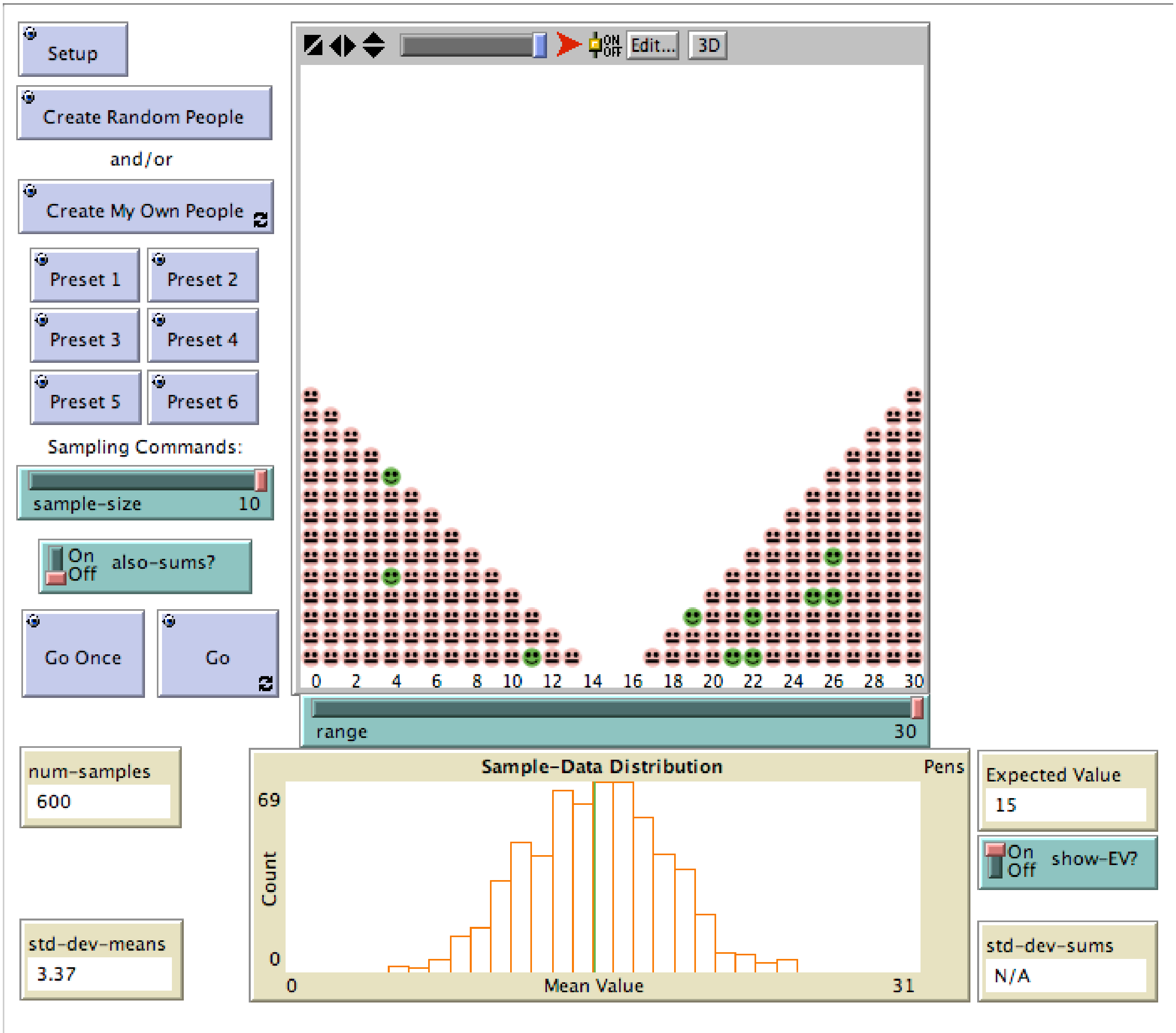
Prove: $\mathbb{E}\bar{x}_n = \mu$
$= \mathbb{E}[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)]$
$= \frac{1}{n}[\mathbb{E}x_1 + \mathbb{E}x_2 + \cdots + \mathbb{E}x_n]$
$= \frac{1}{n}[\mu + \mu + \cdots + \mu] = \frac{1}{n}[n\mu] = \mu$

Prove: $\mathbb{D}\bar{x}_n = \frac{1}{n}\sigma^2$
$= \mathbb{D}[\frac{1}{n}(x_1 + x_2 + \cdots + x_n)]$
$= (\frac{1}{n})^2[\mathbb{D}x_1 + \mathbb{D}x_2 + \cdots + \mathbb{D}x_n]$
$= (\frac{1}{n})^2[\sigma^2 + \sigma^2 + \cdots \sigma^2] = \frac{1}{n}\sigma^2$

NetLogo Demo / Central Limit Theorem sample distr.

Distrib of sample mean: → Normal distrib
For any underlying distrib.

Normal Distrib: $N(\mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$
(Gaussian) w/ $\mu = \mathbb{E}\xi$, $\sigma^2 = \mathbb{D}\xi$


Setup
Create Random People
and/or
Create My Own People
Preset 1
Preset 2
Preset 3
Preset 4
Preset 5
Preset 6
Sampling Commands:
sample-size 10
On Off also-sums?
Go Once
Go
Edit...
3D
0 2 4 6 8 10 12 14 16 18 20 22 24 26 28 30
range 30
num-samples
600
std-dev-means
3.37
Sample-Data Distribution
Pens
69
Count
0
0
Mean Value
31
Expected Value
15
On Off show-EV?
std-dev-sums
N/A

SySc 512, Session 17, Stat Inf (cont.)

Central Limit Theorem

("Distrib of $\bar{X}_n \to N(\mu,\sigma)$ as $n \to \infty$")

Or, more precisely:

Let $Z_n = \frac{n\bar{X}_n - n\mu}{\sigma\sqrt{n}}$ (standardized)

If $\Phi(z)$ is the cumulative distrib of $N(0,1)$
$\Rightarrow \lim_{n\to\infty} P\{Z_n \le z\} = \Phi(z)$

Proof:

1st we need: Characteristic function

$$f_\xi(t) = \mathbb{E}e^{i\xi t}, \quad t \in \mathbb{R}$$

Discrete: $f_\xi(t) = \sum_{k=0}^{\infty} P_\xi(k)e^{ikt}$

Continuous: $f_\xi(t) = \int_{-\infty}^{\infty} p_\xi(x)e^{ixt}\,dx$

($f_\xi(t)$ is a Fourier transform of the distrib)

Taylor expand exponential:

$$f_\xi(t) = \mathbb{E}e^{i\xi t} = \mathbb{E}(1 + i\xi t - \tfrac{1}{2}\xi^2 t^2 + o(t^3))$$
$$= 1 + it\mu - \tfrac{1}{2}t^2(\sigma^2+\mu^2) + o(t^3)$$

used: $D\xi = \mathbb{E}\xi^2 - \mu^2 = \sigma^2$

(Aside: $\mu = -i\frac{d}{dt}f_\xi(t)\big|_{t=0}$, $\sigma^2 = -\frac{d^2}{dt^2}f_\xi(t)\big|_{t=0} + \mu^2$)

For $N(0,1)$: $f_\xi(t) = 1 - \frac{1}{2}t^2 + o(t^3)$

Let $Y_i = \frac{1}{\sigma}(x_i - \mu)$

$\Rightarrow Z_n = \sum_{i=1}^{n} \frac{Y_i}{\sqrt{n}}$

$$f_{Z_n}(t) = \mathbb{E}\left(e^{i(Y_1+Y_2+\cdots+Y_n)t/\sqrt{n}}\right)$$
$$= \mathbb{E}\left(e^{iY_1t/\sqrt{n}}\, e^{iY_2t/\sqrt{n}} \cdots e^{iY_nt/\sqrt{n}}\right)$$
$$= \mathbb{E}e^{iY_1t/\sqrt{n}} \cdots \mathbb{E}e^{iY_nt/\sqrt{n}}$$
$$= \left[f_Y(t/\sqrt{n})\right]^n = \left[1 - \frac{t^2}{2n} + o(t^3/n)\right]^n \to e^{-t^2/2} \text{ as } n\to\infty$$

Same characteristic as $N(0,1)$

~~Wky~~ Continuity Thm: Convergence of characteristic function $\Rightarrow$ convergence of distribution.

SySc 512, Session 17, Stat Inf (cont.)

Statistic: Value calculated from sample to characterize the population (mean, variance)

Purpose of statistics: Determine your confidence that your statistic accurately characterizes the population.

Key: choose a statistic with a known distribution (indep. of unknown population distrib)

Large-n:

$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ has a normal distrib: $\mathcal{N}(0,1)$

$\frac{(n-1)\bar{S}^2_{n-1}}{\sigma^2}$ has a Chi-squared distrib: χ^2_{n-1}

w/ $\chi^2_n = \frac{1}{\Gamma(\frac{n}{2})} \left(\frac{1}{2}\right)^{\frac{n}{2}} y^{\frac{n}{2}-1} e^{\frac{n}{2}-1} e^{-y/2}$

$\Gamma(n) = n!$, $\Gamma(\alpha) = \int_0^\infty x^{\alpha-1} e^{-x} dx$

Small-n (n < 30)

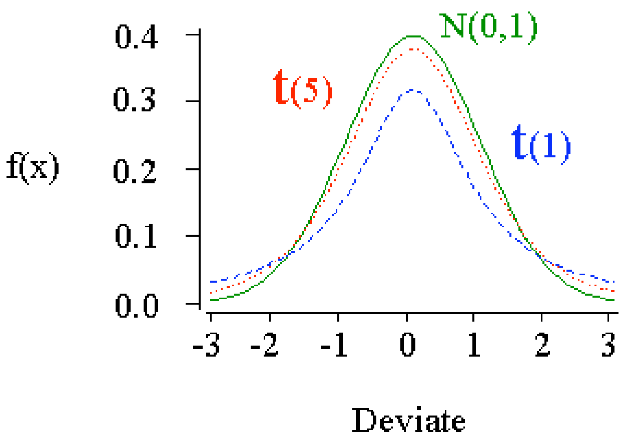
Student's t-distr (Gossett)

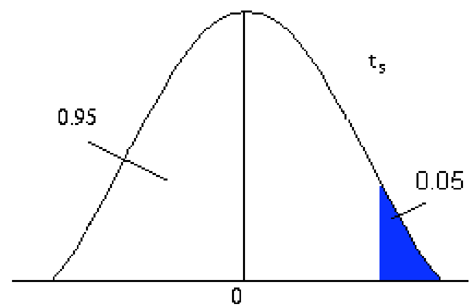
$\frac{\bar{X}_n - \mu}{\bar{S}_n/\sqrt{n}} = \frac{(\bar{X}_n - \mu)/(\sigma/\sqrt{n})}{\sqrt{(n-1)\bar{S}_n^2/(n-1)\sigma}}$; Distrib: $\frac{\mathcal{N}(0,1)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$

t-distrib $\rightarrow \mathcal{N}(0,1)$ for large n

[Figure: slide]

Student's t-distribution





P(t5 > 2.015) = 0.05
P(t20 > 1.725) = 0.05
P(t50 > 1.676) = 0.05
P(tinf > 1.645) = 0.05

Normal distr: P(Z > 1.645) = 0.05 where Z ~ N(0, 1)

SySc 512, Session 17, Stat Inf (cont.)

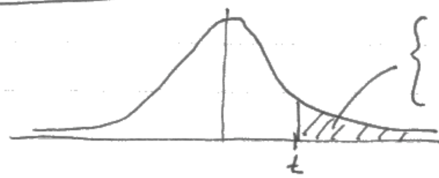
Hypoth Testing

1) Choose null hypothesis: H_0
("not different") [ex: $\mathbb{E}\,\bar{x}_n = \mu$]

2) Choose alternative hypothesis: H_1
("something else") [ex: $\mathbb{E}\,\bar{x}_n > \mu$]

3) Choose significance level (confidence level)
α = % risk of rejecting true Ho.
[ex: $\alpha = 0.01$]

4) Calculate statistic:
[ex: Student's t-test: $t_n = (\bar{x}_n - \mu)/(s_n/\sqrt{n})$]



{Rejection region (R)
$\alpha = \mathbb{P}\{t_n \in R\}$}

5) Get t-value from t-distrib w/ n-1 degrees of freedom:
Let $\Phi_t(\nu) \equiv$ CDF of t-Distrib with ν deg. of f.
$\Rightarrow R = \{t_n \mid \Phi_t(\nu) > 1-\alpha\}$

6) Reject Ho if
$t_n > \Phi_t^{-1}(1-\alpha, n-1)$

[Reported: p-value (observed significance level)
-Smallest fixed level at which Ho is rejected
$p = \mathbb{P}\{t_n > t\} = 1 - \mathbb{P}\{t_n < t\} = 1 - \Phi_t(t_n, \nu)$

7) Interpretation: $\mathbb{E}\,\bar{x}_n$ is not significantly different from μ for signif level α.

[Matlab Demo] hypothTest.m w/ statbox