08/23/19

Lecture 12

AMS 131

Hypergeometric: a finite population has A elements of type 1 and B of type 2, total size (A+B)

you choose n elements random <u>without replacement</u> (SRS)

$X = \begin{pmatrix} \#\text{ of type 1} \\ \text{in sample} \end{pmatrix} \sim$ hypergeometric (A, B, n)

$$f_X(x \mid A, B, n) = \underbrace{\frac{\binom{A}{x}\binom{B}{n-x}}{\binom{A+B}{n}}}_{\text{PMF}} \underbrace{I[\max\{0, n-B\} \le x \le \min(n, A)]}_{\text{support}}$$

(A, B, n) non negative $A+B \ge n$

1. $E(X) = n \cdot \frac{A}{A+B}$ 2. $V(X) = n \cdot \left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\left(\frac{A+B-n}{A-B-1}\right)$

If taken <u>with replacement</u> (IID), X would be Binomial with same n and:

$p = \frac{A}{A+B}$ $E(X) = np = n \cdot \frac{A}{A+B}$ $V(X) = np(1-p) = n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$

If $T = A+B$, total # of elements in population

	mean	variance (of X)
with replacement (IID)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)$
without replacement (SRS)	$n\left(\frac{A}{A+B}\right)$	$n\left(\frac{A}{A+B}\right)\left(\frac{B}{A+B}\right)\underline{\left(\frac{T-n}{T-1}\right)}$

SRS more information, no repetition, smaller variance

$0 \le \alpha = \frac{T-n}{T-1} \le 1$ called the finite population correction.

a.) $n=1$, $\alpha=1$ $\leftrightarrow$ SRS = IID (only 1 element)

b.) $n=T$, $\alpha=0$ $\leftrightarrow$ exhaust the entire population with SRS, no uncertainty left

c.) fix n, allow $T \uparrow$, $\alpha \to 1$ $\leftrightarrow$ small sample from a large population, SRS $\doteq$ IID

AMS 131 Lecture 12 (cont.)

$X \sim \text{Poisson}(\lambda) \quad \lambda > 0$

PMF $f_X(x) = \frac{\lambda^x e^{-\lambda}}{x!} I_{\{0,1,\ldots\}}(x)$ ← support of x

$E(X) = \lambda$, $V(X) = \lambda$] $\frac{V(X)}{E(X)} = 1$ variance-mean ratio (VTMR)

$\psi_X(t) = e^{\lambda(e^t - 1)}$ unrealistic, most R.V.s that represent counts of occurrences of events in fixed time interval lengths = VTMR > 1

The Poisson and Binomial both count the number of "successes" as time goes on, so:

when (n is large, p is close to 0) → $\text{Binomial}(n,p) \doteq \text{Poisson}(n \cdot p)$

n pos. integer $X \sim \text{Binomial}(n,p)$

$0 < p < 1$ $Y \sim \text{Poisson}(\lambda) \quad \lambda > 0$

any sequence of numbers $\{p_n\}_{n=1}^{\infty}$ of values between 0 and 1 with $\lim_{n\to\infty} n \cdot p_n = \lambda$

n = # of successes.

PMF Binomial → PMF Poisson

$f_X(x \mid n, p_n) \longrightarrow n\to\infty \longrightarrow f_Y(y \mid \lambda)$

A Poisson Process with rate λ per unit time is a stochastic process with two properties

1. # of arrivals in every interval of time length t $\sim$ Poisson (λt)
2. # of arrivals in all disjoint (non overlapping) time intervals are independent.

Cryptosporidium case study

people get sick when $\frac{1 \text{ oocyst}}{5 \text{ liters}}$ ← organism, at one stage in life

Concentration is ← 1.3 gal of water

Suppose oocysts occur according to a Poisson process with rate λ oocysts per liter. A filtering system can capture all of the oocysts in a sample, but only has a probability p of detecting if they are actually there.

Y = # of oocysts in water (actual) t liters

$X_i = \begin{cases} 1 & \text{if oocyst } i \text{ gets counted} \\ 0 & \text{else} \end{cases}$ (counting events are independent)

X = # of counted oocysts $\quad (X \mid Y=y) = \sum_{i=1}^{y} X_i \sim \text{Binomial}(y, p)$

X only becomes binomial given $Y=y$

The Law of Total Probability says: $f_X(x) = P(X=x) = \sum_{y=0}^{\infty} P(Y=y)P(X=x \mid Y=y)$ ← for all $x=0,1\ldots$

↳ $P(Y=y) = \frac{(\lambda t)^y e^{-\lambda t}}{y!}$ for $y=0,1\ldots$ and $P(X=x \mid Y=y) = \binom{y}{x} p^x (1-p)^{y-x}$

if $X=x$, then $Y \geq x$ – at least as large as the number of oocysts detected

$$f_X(x) = \sum_{y=x}^{\infty} \binom{y}{x} p^x (1-p)^{y-x} \frac{(\lambda t)^y e^{-\lambda t}}{y!} = \frac{e^{-p\lambda t}(p\lambda t)^x}{x!}$$

$X \sim \text{Poisson}(p\lambda t)$: losing $(1-p)$ proportion to faulty counting lowers the rate of the Poisson process from λ/liter to $\lambda \cdot p$/liter

Poisson mixture of Binomials (= another Poisson)

In practice, p is small (close to zero). How much water (t liters) do you need to filter to achieve: $P(\text{at least 1 oocyst detected}) \geq 1-\alpha$ for small α?

AMS 131

Lecture 12 (cont.)

$P(\text{at least 1 detected}) = 1 - P(\text{none detected})$

$= 1 - P(X=0) = 1 - e^{-p\lambda t} \geq 1-\alpha$

$\leftrightarrow \alpha \geq e^{-p\lambda t} \leftrightarrow \ln(\alpha) \geq -p\lambda t \rightarrow t \geq \frac{-\ln(\alpha)}{p\lambda}$

$\alpha = 0.01$ (failure rate) $\quad p = 0.1$

$\lambda = 0.2$ / liter $= 1$ / 5 liters

to achieve 99%, t has to be at least 230.3 liters

Negative Binomial Distribution: you're watching a potentially endless sequence of Bernoulli trials with constant success probability p.

X = # of failures before r^{TH} success $\quad r \geq 1$ integer

$X \sim$ Negative Binomial dist. PMF: $f_X(x|r,p) = \binom{r+x-1}{x} p^r (1-p)^x \cdot I_{\{0,1,2\ldots\}}(x)$

now we're counting failures instead of successes, there are two ways to estimate (unknown) p.

1. decide ahead of time to sample n success/failures and record the (random) # S of success you see, a good estimate would be $\hat{p}_B = \frac{S}{n}$ (binomial)
2. Sample until you see s successes, then record the # of trials N needed to accumulate that many successes, estimate $\hat{p}_{NB} = \frac{s}{N}$ (neg. binomial)

Binomial: numerator random, denominator fixed

N-Binomial: numerator fixed, denominator random.

Set $r=1$ and record the number X of failures until the first success, X follows the Geometric(p) dist.

PMF $f_X(x|p) = p(1-p)^x I_{\{0,1\ldots\}}(x)$ ← parameter p, support of X

$X_1, \ldots, X_n$ IID Geometric(p)

$\rightarrow \sum_{i=1}^{n} X_i \sim$ Negative Binomial(n,p)

$X_1 \ldots X_n$ IID Bernoulli(p)

$\rightarrow \sum_{i=1}^{n} X_i \sim$ Binomial(n,p)

$X \sim$ Negative Binomial(r,p)

$$M_X(t) = \left[\frac{p}{1-(1-p)e^t}\right]^r \text{ for } t < \log\left(\frac{1}{1-p}\right)$$

$$E(X) = \frac{r(1-p)}{p} \qquad V(X) = \frac{r(1-p)}{p^2}$$

$X \sim$ Geometric(p)

$P(X = k+t \mid X \geq k) = P(X = t)$ $\left\{\begin{matrix} k \\ t \end{matrix}\right\}$ non negative integers

Wait k trials, whats the chance you succeed on the next $k+t$ trial, just $P(X=t)$ waiting t trials

- memoryless property of the Geometric distribution
  forget the condition of waiting k trials
  ONLY discrete distribution with this property.

F F F F F S

X = # of failures until first success

Y = # of failures, starting at trial $k+1$ until next success

Y has the same dist. as X and is independent of what happened in the first k trials

↳ "the process has no memory"

AMS 131 | Lecture 12 (cont.)

Continuous distributions

Normal (Gaussian) dist $X \sim \text{Normal}(\mu, \sigma^2)$

μ = mean, σ^2 variance $0 < \sigma^2 < \infty$

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \quad f_X(x \mid \mu, \sigma^2)$$

converting to standard units

$\sigma \geq 0$ $\sigma = 0$  , use $\sigma > 0$

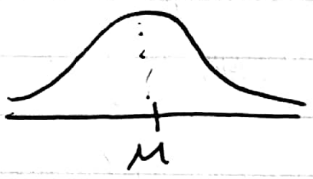
$-\infty < \mu < \infty$

1. Many random processes have this dist. shape
2. The Central Limit Theorem

$X \sim \text{Normal}(\mu, \sigma^2) = N(\mu, \sigma^2)$

$E(X) = \mu$, $V(X) = \sigma^2$, $SD(X) = \sigma$

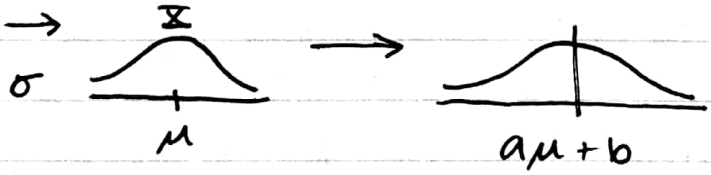
$$\psi_X(t) = \exp\left(\mu t + \frac{\sigma^2 t^2}{2}\right)$$



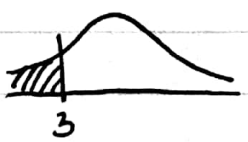
Center of symmetry = median = mean = mode = μ

$(X \mid \mu, \sigma^2) \sim N(\mu, \sigma^2)$

$Y = aX + b$, $a \neq 0$, b constants



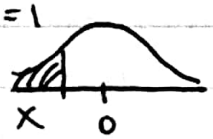
Normality is preserved under linear transformation



$X \sim N(\mu, \sigma^2)$

$P(X \leq 3)$ CDF of Normal has no closed form.

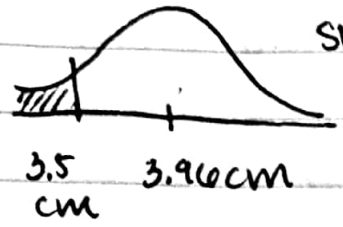
SD = 1



Standard Normal Dist. $\mu = 0$, SD = 1

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{t^2}{2}\right] dt$$

n = 103 Monarch butterflies, wing length.



$$\begin{bmatrix} Y_1 = 4.1 \\ Y_2 = 3.3 \\ \cdots \\ Y_n = 4.7 \end{bmatrix} \quad n = 103$$

$\bar{Y} = 3.96$ cm

$\sigma_Y = 0.29$ cm

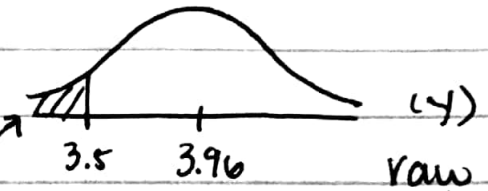
$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \qquad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$P(\bar{Y} \leq 3.5)$?

1. $\begin{bmatrix} Y_1 = 4.1 \\ Y_2 = 3.3 \\ \cdots \\ Y_n = 4.7 \end{bmatrix} \rightarrow \begin{bmatrix} 0 \\ 1 \\ \cdots \\ 0 \end{bmatrix}$

$$X = \begin{cases} 1 & \text{if } Y \leq 3.5 \\ 0 & \text{else} \end{cases}$$

$\mu = \frac{\text{SUM}}{103}$

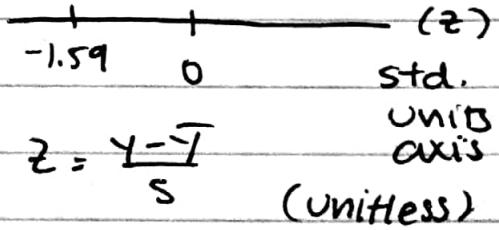


2. $\begin{bmatrix} 3.2 \\ 3.3 \\ \cdots \\ 3.5 \\ 3.5 \\ 3.5 \\ 3.6 \\ \cdots \end{bmatrix}$ 8 ↓ 103 ↓

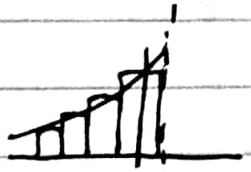
(exact) $= \frac{8}{103} \doteq 7.8\%$

Sort, then count

online calc = 5.63%

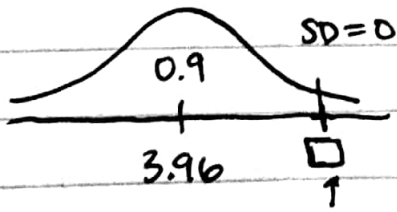


$Z = \frac{Y - \bar{Y}}{S}$



continuity correction

0.0787 = 7.87%



What is the 90TH percentile of the wing length dist?

- Inverse CDF problem

online calc = 4.332

PDF of $\bar{Y} \sim$ Normal (0,1) CDF

$$\phi_{\bar{Y}}(x) \triangleq \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \rightarrow \Phi_{\bar{Y}}(x) \triangleq \int_{-\infty}^{x} \phi_{\bar{Y}}(t)\,dt$$

lower case

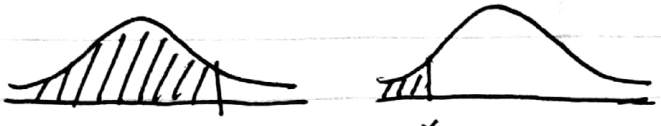
AMS 131 | Lecture 12 (cont.)

1. e^{-cx^2} $c>0$ has no anti-derivative in closed form, we have approximations in a table using numerical integration

2. The Normal PDF is symmetric (for all $X \in \mathbb{R}$)

$\Phi(-x) = 1 - \Phi(x)$ $\quad \phi_X(x) = \phi_X(-x)$ ← same height

$\Phi^{-1}(p) = -\Phi^{-1}(1-p)$ for all $0<p<1$



Symmetric on both sides

$\Phi_X(x) + \Phi_X(-x) = 1$

$X \sim \text{Normal}(\mu, \sigma^2) \rightarrow \frac{X-\mu}{\sigma} \sim N(0,1)$ standard

$$F_X(x) = \Phi\left(\frac{x-\mu}{\sigma}\right)$$

$$F_X^{-1}(p) = \mu + \sigma\Phi^{-1}(p)$$

Empirical Rule
- $(\mu \pm \sigma)$ = 68% of data.
- $(\mu \pm 2\sigma)$ = 95%.
- $(\mu \pm 3\sigma)$ = 99.7%.

} exact for all normal distributions

$X_1 \dots X_k$ independent $\quad X_i \sim \text{Normal}(\mu_i, \sigma_i^2)$

$$\rightarrow \sum_{i=1}^{k} X_i \sim \text{Normal}\left(\sum_{i=1}^{k}\mu_i, \sum_{i=1}^{k}\sigma_i^2\right)$$ ←

additive property, why Normals are indexed by $V(X)$

women height	men height
$N(65\text{in}, (3.2)^2\text{in}^2)$	$N(69.5\text{in}, (3.3)^2\text{in}^2)$
μ $\quad \sigma = 3.2\text{in}$	μ $\quad \sigma = 3.3\text{in}$

1 woman with height $\boxtimes$
1 man with height $\boxtimes$
} chosen randomly, independent

$P(\text{woman taller than man}) = P(\boxtimes \geq \boxtimes)$ due to D:

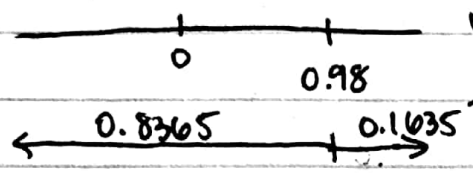
$D = \boxtimes - \boxtimes$ looking for positive D ↙ linear operation

$D \sim N(65 - 69.5 = -4.5\text{in}, (3.2)^2 + (3.3)^2 = 21.1\ \text{in}^2)$

$SD = \sqrt{21.1} = 4.6\text{in}$

convert to standard units:

$$\frac{0 - (-4.5)}{4.6} = +0.98$$



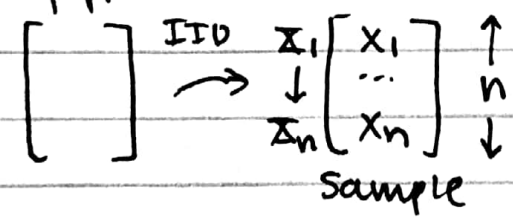
using online calc.

→ so $P(\boxtimes \geq \boxtimes) \doteq 16\%$

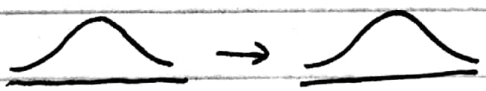
about 1 in 6

R.V. $\boxtimes_1, \ldots, \boxtimes_n$ → sample mean of $(\boxtimes_1, \ldots, \boxtimes_n)$

is $\bar{\boxtimes}_n = \frac{1}{n}\sum_{i=1}^{n} \boxtimes_i$ $\begin{cases} \boxtimes_i \overset{IID}{\Rightarrow} N(\mu, \sigma^2) \\ (i = 1, \ldots, n) \end{cases}$



$\bar{\boxtimes}_n = \frac{1}{n}\sum_{i=1}^{n} \boxtimes_i = (\bar{x})$ linear transformation



$$E(\bar{\boxtimes}_n) = E\left(\frac{1}{n}\sum_{i=1}^{n} \boxtimes_i\right)$$

$E(\bar{\boxtimes}_n) = \mu \leftrightarrow \bar{\boxtimes}_n$ is unbiased for μ.

$\bar{\boxtimes}_n$ is an unbiased estimator

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} \boxtimes_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(\boxtimes_i)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \mu = \mu \checkmark$$

In frequentist statistics, the standard deviation on an estimator $\hat{\theta}$ (R.V.) of a parameter θ is called the standard error $SE(\hat{\theta})$ of $\hat{\theta}_k$

AMS 131 Lecture 12 (cont.)

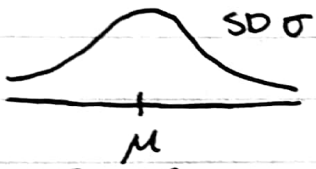
$SD(\bar{X}_n) \triangleq SE(\bar{X}_n) = \frac{c}{f(n)}$

$SD\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n} SD\left(\sum_{i=1}^{n} X_i\right)$ - hard?

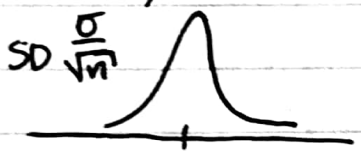
$V(\bar{X}_n) = V\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2} V\left(\sum_{i=1}^{n} X_i\right) \overset{\text{(I)}}{=} \frac{1}{n^2}\sum_{i=1}^{n} V(X_i)$

$\overset{\text{ID}}{=} \frac{1}{n^2}\sum_{i=1}^{n} \sigma^2 = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$ ✱

$SD(\bar{X}_n) = SE(\bar{X}_n) = \sqrt{V(\bar{X}_n)} = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$ square root law



PDF of $\bar{X}_1$ (population PDF)



PDF of $\bar{X}_n$ $(n > 1)$

get closer with $n \to \infty$