FISH 604
Module 2 (part 1) :
Basic statistical concepts

Instructor: Franz Mueter

Lena Point, Rm 315

796-5448

fmueter@alaska.edu

Today's goals

You should understand…

- … difference between arithmetic, geometric, and harmonic mean
- … standard error, standard deviation, and coefficient of variation

You should be able to…

- … estimate location and spread for any random variable
- … quickly summarize data statistically & graphically
- … estimate variance of the sum, the product, and the ratio of random variables

MSL FISH 604

Basic statistical concepts

- Data summaries
  - Location: Mean, median, quantiles
  - Spread: Variance, Standard deviation, MAD, IQR
  - Graphical summaries (Histograms, dotplots, boxplots)
- Probability and probability distributions
- Distributions
  - Discrete (Binomial, Multinomial, Poisson)
  - Continuous (Uniform, Exponential, Normal or Gaussian, Log-normal, Gamma)

MSL FISH 604

Data summaries

- Data: $x_1, x_2, \ldots x_n$
- Measures of location / central tendency:
  - **Mean (sample mean)**

Arithmetic mean: $$AM_x = \bar{x} = \sum_{i=1}^{n} x_i \Big/ n$$

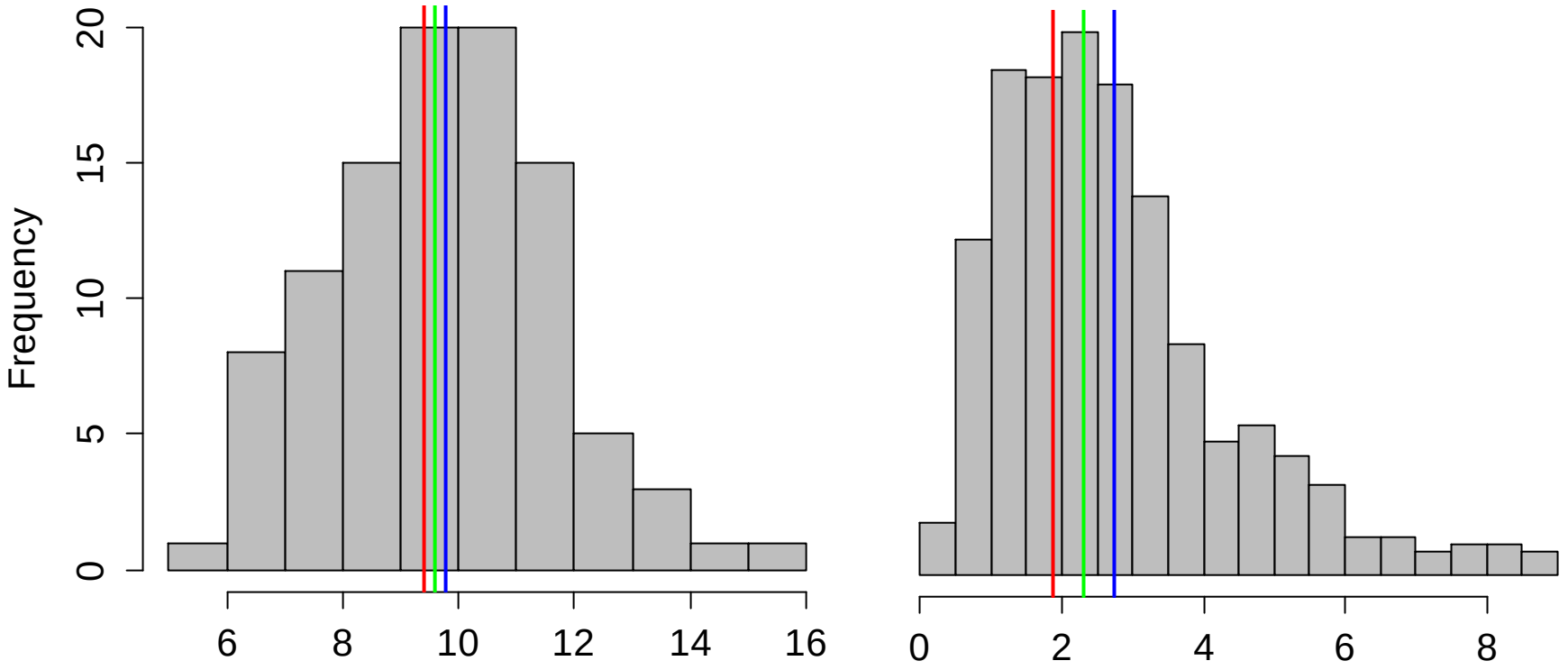
Geometric mean: $$GM_x = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log(x_i)\right) = \sqrt[n]{x_1 x_2 \ldots x_n}$$

Harmonic mean: $$HM_x = \frac{1}{\frac{1}{n}\sum_{i=1}^{n}\frac{1}{x_i}}$$

$$HM \leq GM \leq AM$$

(always!)

Examples: Arithmetic, geometric, and harmonic means



→ See script *Mod 2(1).R* to explore other examples

MSL FISH 604

Data summaries

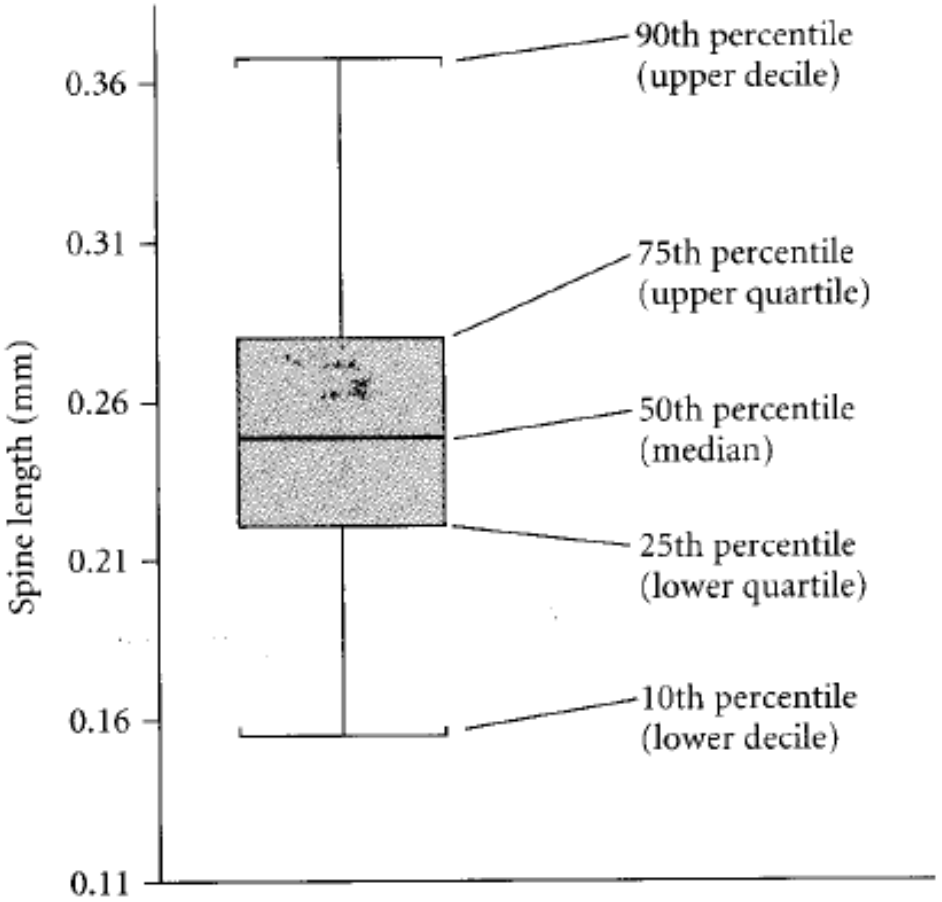
- Data: $x_1, x_2, \ldots x_n$
- Other measures of location:
  - Median (robust measure)
    - $Pr(x_i < Median) = 0.5$
    - 50% of *data points* < Median < 50% of data points
  - Quantiles
    - Quartiles
    - Quantiles corresponding to any probability

→ See R script '*Mod 2(1).R*'

Boxplots

MSL FISH 604

Figure 3.6 Box plot illustrating quantiles of data from Table 3.1 ($n = 50$). The line indicates the 50th percentile (median), and the box encompasses 50% of the data, from the 25th to the 75th percentile. The vertical lines extend from the 10th to the 90th percentile.



From Ch.3 Gotelli & Ellison

Data summaries

MSL FISH 604

Measures of spread / dispersion

- Variance / Standard deviation

- Coefficient of Variation:

MSL FISH 604

Data summaries

Measures of spread / dispersion

- Variance / Standard deviation

$$Var(x) = \sum_{i=1}^{n} (x_i - \bar{x})^2 \Big/ (n-1)$$

$$SD(x) = s = \sqrt{Var(x)}$$

- Coefficient of Variation: $CV(x) = s / \bar{x}$

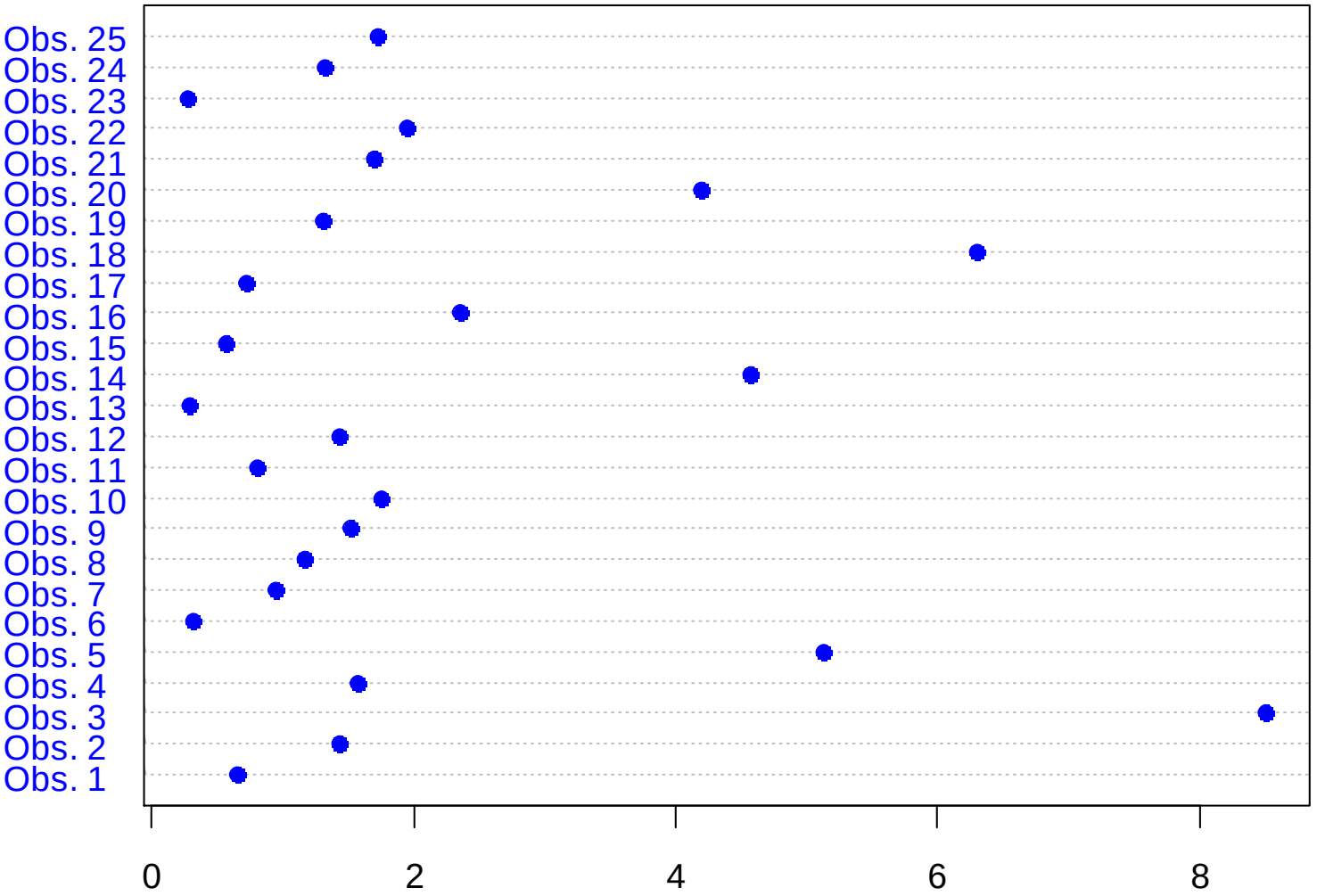
- Median Absolute Deviation (robust measure)

$$MAD(x) = 1.4826 \cdot Median\left|(\mathbf{x} - \bar{x})\right|$$

- Range, interquartile range

Graphical data summaries

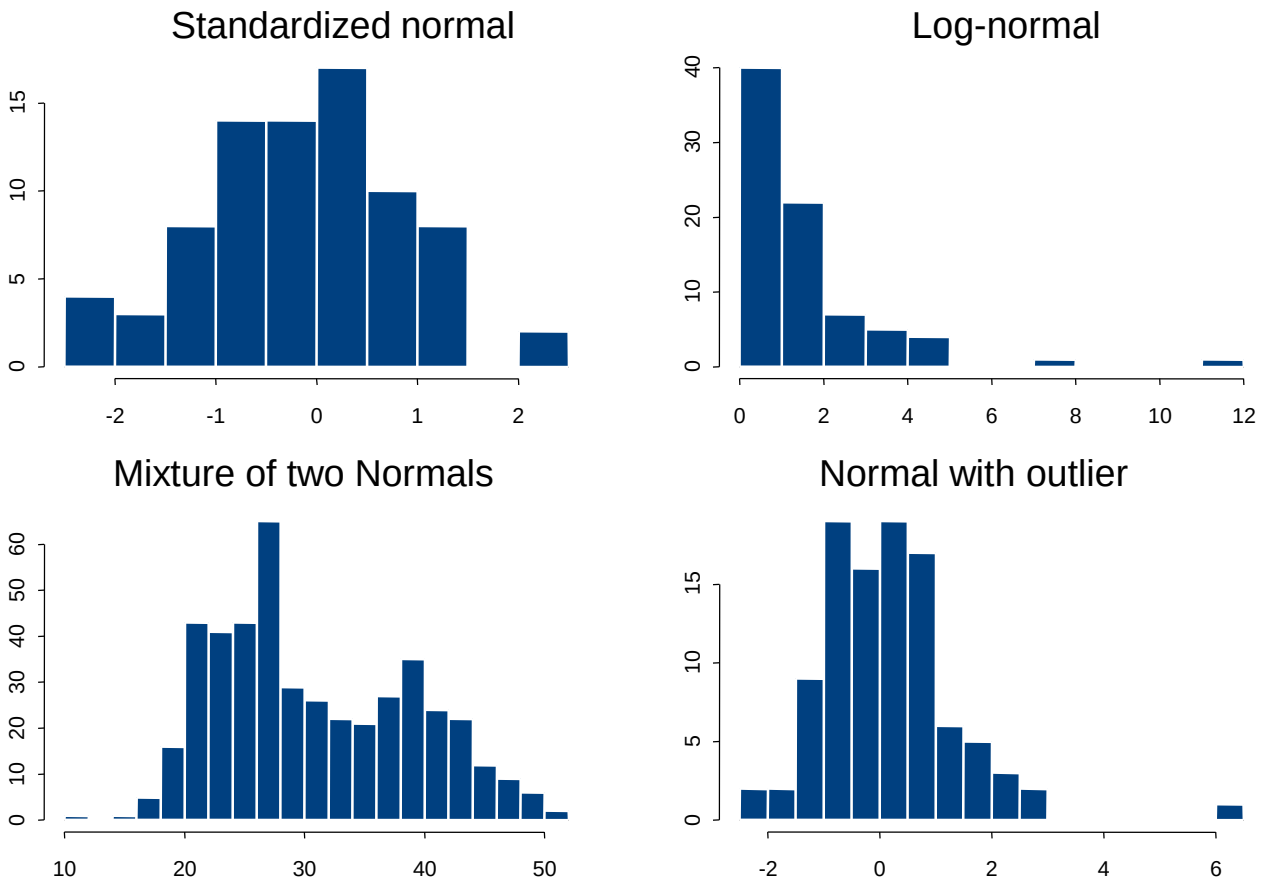
MSL FISH 604



MSL FISH 604

Graphical data summaries

Histograms: Simulated data

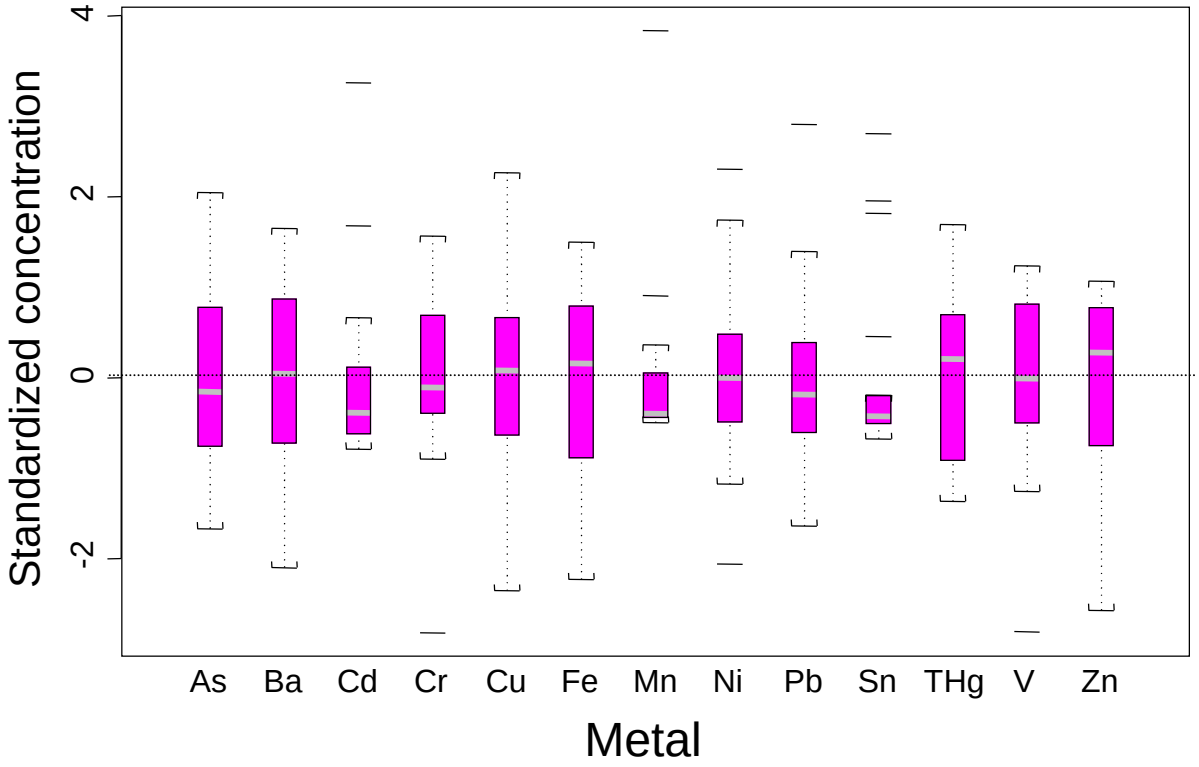


MSL FISH 604

Graphical data summaries

Boxplots

Standardized concentrations of 13 metals in Beaufort Lagoon sediments



Expectation and Variance

MSL FISH 604

Expectation or expected value (mean)

- Formal definition:

(discrete)

(continuous)

- Some useful theorems:

(assuming independence)

MSL FISH 604

Expectation and Variance

Expectation or expected value (mean)

- Formal definition:

$$E(x) = x_1 P(X = x_1) + \ldots + x_n P(X = x_n) = \sum_{j=1}^{n} x_j P(X = x_j)$$ **(discrete)**

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx$$ **(continuous)**

- Some useful theorems:

$$E(cX) = cE(X)$$

$$E(X + Y) = E(X) + E(Y)$$

$$E(XY) = E(X)E(Y)$$ **(assuming independence)**

Example (part 1: Expected value)

What was the total catch C (in pounds) of age-3 pollock caught in the Eastern Bering Sea in 2011?

- Estimated catch at age-3 in numbers from stock assessment (Ianelli et al. 2012):

$$E(N_3) = 193.2 \text{ million fish}$$

with an assumed error of 5% ($CV = 0.05$)

- Mean weight of age-3 fish from random sampling was $E(w_3) = 0.290$ kg with a sampling error of 2% ($CV = 0.02$)
- 1 kg = 2.21 pounds (conversion factor $f = 2.21$ lb/kg)

→ Compute expected value of catch:

$$E(C) = E(f * w_3 * N_3)$$

Expected value of catch:

$$E(C) = E(f * w_3 * N_3)$$
$$= ??$$

'f' is a constant!

'w_3', 'N_3' are independent!

MSL FISH 604

Variance

- Quantifies "spread" or uncertainty
- Definition:
- Some useful theorems:

Variance

- Quantifies "spread" or uncertainty
- Definition: $\mathrm{Var}(X) = \sigma^2 = E[(X-\mu)^2]$
- Some useful theorems:

$$\mathrm{Var}(X) = E\left[(X-\mu)^2\right] = E\left(X^2\right) - \mu^2$$

$$\mathrm{Var}(cX) = c^2 \cdot \mathrm{Var}(X)$$

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \underbrace{2\mathrm{Cov}(X, Y)}$$

=0 if X,Y uncorrelated

MSL FISH 604

Covariance

- Definition:

- Useful theorems:

MSL FISH 604

Covariance

- Definition:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= E\left[\left(X-\mu_X\right)\cdot\left(Y-\mu_Y\right)\right] \\ &= E(XY)-E(X)E(Y)\end{aligned}$$

- Useful theorems:

$$\mathrm{Cov}(aX,bY) = ab\cdot\mathrm{Cov}(X,Y)$$

$$\mathrm{Cov}(X,Y+Z) = \mathrm{Cov}(X,Y)+\mathrm{Cov}(X,Z)$$

MSL FISH 604

Correlation

- Definition:

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}$$

(dimensionless)

$-1 < \rho < 1$

Variance of a product

- General law for variance of a product:

$$\mathrm{var}(x_1 x_2) \approx \mu_2^2 \mathrm{var}(x_1) + \mu_1^2 \mathrm{var}(x_2) + 2\,\mathrm{cov}(x_1 x_2)\mu_1 \mu_2$$

where: $\mu_i = \mathrm{E}(x_i)$

- Two <u>independent</u> random variables:

$$\mathrm{var}(x_1 x_2) = \mu_2^2 \mathrm{var}(x_1) + \mu_1^2 \mathrm{var}(x_2) + \mathrm{var}(x_1)\mathrm{var}(x_2)$$

Example (part 2: Variance)

MSL FISH 604

What is the variance of the total catch of age-3 pollock caught in the Eastern Bering Sea in 2011?

- Estimated catch at age-3 in numbers from stock assessment (Ianelli et al. 2012):

  $E(N_3) = 193.2$ million fish

  with an assumed error of 5% ($CV = 0.05$)
- Mean weight of age-3 fish from random sampling was $E(w_3) = 0.290$ kg with a sampling error of 2% ($CV = 0.02$)
- 1 pound = 0.453 kg (conversion factor c = 0.453)

→ Compute variance of catch:

$$var(C) = var(c * w_3 * N_3)$$

Variance, catch of age-3 pollock:

$$\mathrm{var}(C) = \mathrm{var}(f * w_3 * N_3)$$

$$= ??$$

Variance, catch of age-3 pollock:

$$\begin{aligned} \mathrm{var}(C) &= \mathrm{var}(f \cdot w_3 \cdot N_3) \\ &= ?? \\ &= f^2 \cdot \mathrm{var}(w_3 \cdot N_3) \\ &= f^2 \cdot \{E(N_3)^2 \cdot \mathrm{var}(w_3) + E(w_3)^2 \cdot \mathrm{var}(N_3) \\ &\qquad + \mathrm{var}(w_3) \cdot \mathrm{var}(N_3)\} \end{aligned}$$

Given: $CV(N_3) = sd(N_3)/E(N_3) = 0.05$

hence: $sd(N_3) = 0.05 \cdot E(N_3)$

and: $\mathrm{var}(N_3) = sd(N_3)^2 = 0.05^2 \cdot E(N_3)^2$

similarly: $\mathrm{var}(w_3) = 0.02^2 \cdot E(w_3)^2$

Substituting these expressions for $\mathrm{var}(N_3)$ and $\mathrm{var}(w_3)$ above yields: $\mathrm{var}(C) = 44.28$

$sd(C) = 6.65$

and a 95% CI is given by: {180.2, 206.2}

MSL FISH 604

Variance estimation

- If no analytical solution can be obtained for the variance of a random variable, we can use:
  - Delta method (Approximation based on Taylor series expansion)
  - Jacknife (largely superceded by bootstrap)
  - Bootstrap (computer intensive method based on re-sampling)

Delta method

MSL FISH 604

Approximate method to estimate expected value and variance of a function of X, if $E(X)=\mu_x$ and $var(X)=\sigma_x^2$ are known

- Let $Y = g(X)$, then:

$$E(Y) = \mu_y \approx g(\mu_x) \qquad \text{(rough approximat ion)}$$

$$E(Y) = \mu_y \approx g(\mu_x) + \tfrac{1}{2}\sigma_x^2 \cdot g''(\mu_x) \qquad \text{(better approximat ion)}$$

$$Var(Y) = \sigma_y^2 \approx \sigma_x^2 [g'(\mu_x)]^2$$

- Example: Approximate mean and variance of the inverse of a variable: $Y = g(X) = 1/X$

MSL FISH 604

Bootstrap estimate of variance

- A method to provide an estimate the variance of a summary statistic (T) from a sample alone, without making any distributional assumptions
- The sample is taken to "represent" the distribution of the whole population
- Many random "bootstrap samples" are drawn from this "population" (with replacement).
- The statistic, T, is calculated from each bootstrap sample and tallied (=bootstrap estimates of T, or T_B)
- The variance of the bootstrap estimates (T_b) provides an estimate of the variance of T (in fact, the full distribution of T_B estimates the full distribution of T)

MSL FISH 604

Exercise (in class)

- Delta method example to compute variance of inverse of a variable
- Simple Bootstrap examples to estimate variance of the median

Review / preview
Basic statistical concepts

MSL FISH 604

- Data summaries
  - Location: Mean, median, quantiles
  - Spread: Variance, Standard deviation, MAD, IQR
  - Graphical summaries
  - Expectation & Variance; Variance estimation
- Probability and probability distributions
- Distributions
  - Discrete (Binomial, Multinomial, Poisson)
  - Continuous (Uniform, Exponential, Normal or Gaussian, Log-normal, Gamma)

Next time

Reading

MSL FISH 604

Reading assignment

- Gotelli, N.J., and Ellison, A.M. 2004. A Primer of Ecological Statistics. **Chapter 3**

Further background (theoretical)

- Casella, G., and Berger, R. 2002. Statistical Inference. Duxbury, Pacific Grove, CA.
- Rice, J.A., 1995. Mathematical Statistics and Data Analysis. Duxbury Press, Belmont, CA.