Statistical Inference

- drawing conclusions about a population parameter, based on a sample estimate.

Population: GRE results for a new exam format on the quantitative section
Sample: n=30 test scores

	Population		Sampling Dist. for $\bar{X}$
shape	normal ?		approx. normal
mean	μ	(unknown)	μ
SD	σ =100 (assume known)		$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{100}{\sqrt{300}} = 5.8$

100(1-α)% ***Normal*** Confidence Interval for μ

$\bar{x} \pm z_{\alpha/2} \bullet \frac{\sigma}{\sqrt{n}}$ where $z_{\alpha/2}$ is the upper tail critical value from the normal distribution

ex: GRE scores (n=30, σ=100, $\bar{x} = 520$)

95% CI for μ: $520 \pm 1.96 \bullet \frac{100}{\sqrt{30}} = 520 \pm 35.78 = (484.22,\ 555.78)$

100(1-α)% ***Student's T*** Confidence Interval for μ

$\bar{x} \pm t_{\alpha/2} \frac{s}{\sqrt{n}}$ where $t_{\alpha/2}$ is the upper tail ***T*** critical value with *n-1* d.f.

Hypothesis Testing

- Null Hypothesis (H_o) vs. Alternative Hypothesis (H_A)
- α = P(Type I Error) → the level of significance (*l.o.s.*)
- β = P(Type II Error)
- *power* = $1 - \beta$

p-value

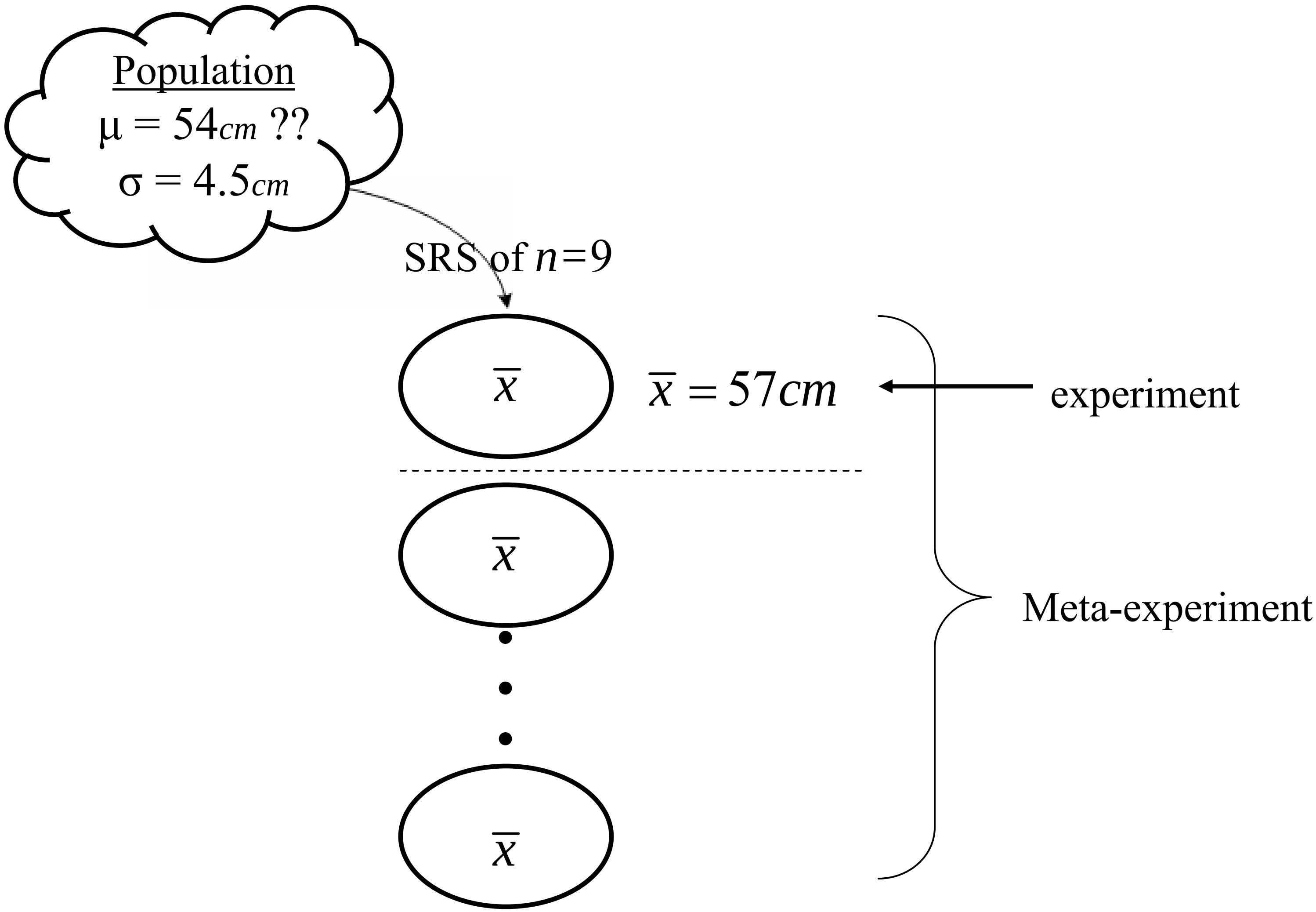
- Measures the strength of the sample evidence against H_o
- Definition:

  The probability, computed assuming that H_o is true, of a sample result ($\bar{X}$) as extreme or more extreme than the one from our sample.

Rule of Thumb for the significance of p-values

- The p-value is the smallest *l.o.s.* α at which we can reject H_o
- If the p-value is less than .05, then our results are statistically significant at the .05 level

Let X be a R.V. denoting the length of a fish in *cm*.



Does $\overline{x} = 57$ give strong evidence that μ>54 *cm*?

How likely are we to get $\overline{x} = 57$ if μ=54 *cm*?

H_o: μ=54 *cm* $\qquad \sigma_{\bar{X}} = \frac{4.5\ cm}{\sqrt{9}} = 1.5\ cm$

H_a: μ>54 (μ_a = 58) $\qquad \alpha = .05$

<u>4 Steps for finding the Power in a test of hypotheses</u>

1) Write the RR for H_o in terms of z-scores:

2) Write the RR for H_o in terms of $\bar{X}$:

3) Find the probability of a Type II error <u>if μ=58</u>

4) Power = 1 – P(Type II Error):

Probability Distributions for Random Variables (RVs)

- Probability Distribution (for a *discrete* RV) – represented graphically as a ***Probability Histogram***
  - Indicates the possible values for a RV
  - Indicates how to assign probabilities for the possible values: $p(x) = P(X = x)$
- Probability Distribution (for a *continuous* RV) – represented as a ***Probability Density Curve***
  - Areas under a smooth curve, $f(x)$, indicate probabilities of values in a given range

Expected Value of a Random Variable

- a weighted average of all possible values for *X*, weighted by the probability of each value
  $E(X) = \mu$, the mean for the RV
  - $E(X) = \sum_{x=-\infty}^{\infty} x\, p(x)$ for a *discrete* RV
  - $E(X) = \int_{x=-\infty}^{\infty} x\, f(x)\, dx$ for a *continuous* RV

Example (*discrete* RV):
Y = # heads in two tosses of a fair coin

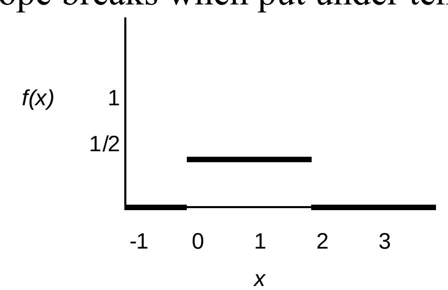
P(Y = 0) = ¼
P(Y = 1) = ½
P(Y = 2) = ¼ and zero otherwise

$$E(Y) = 0 \bullet P(Y=0) + 1 \bullet P(Y=1) + 2 \bullet P(Y=2)$$

$$= 0 \bullet \frac{1}{4} \quad + \quad 1 \bullet \frac{1}{2} \quad + \quad 2 \bullet \frac{1}{4} \quad = \quad 1$$

Example (*continuous* RV):
X = the position at which a two-meter with length of rope breaks when put under tension (assuming every point is equally likely)

f(x)= ½ $0 \le x \le 2$, zero otherwise



$$E(X) = \int_0^2 x\left(\frac{1}{2}\right)dx = \frac{1}{4}x^2\Big|_0^2 = 1 - 0 = 1$$

Properties of Expectation
E(a) =a
*E(aX) =a*E(X)*
E(X+a)= E(X) + a
E(X+Y)= E(X) + E(Y)

*E(XY) = E(X) * E(Y)* if X & Y are independent

Variance of a Random Variable

- a weighted average of *squared deviations from the mean*, $[x - E(x)]^2$

$$Var(X) = E[(X-\mu)^2] = \sigma^2$$

  - $Var(X) = \sum_{x=-\infty}^{\infty} \left(x - E(x)\right)^2 p(x)$ for a *discrete* RV
  - $Var(X) = \int_{x=-\infty}^{\infty} \left(x - E(x)\right)^2 f(x)\,dx$ for a *continuous* RV

$$Var(X) = E[(X-\mu)^2] \quad \Rightarrow \quad E(X^2) - [E(X)]^2 \quad \text{for } \textbf{\textit{any}} \text{ RV}$$

Example:
Y = # heads in two tosses of a fair coin

P(Y = 0) = ¼
P(Y = 1) = ½
P(Y = 2) = ¼ and zero otherwise

$$\mu = E(Y) = 1$$

$$Var(Y) = (0-\mu)^2 \bullet P(Y=0) + (1-\mu)^2 \bullet P(Y=1) + (2-\mu)^2 \bullet P(Y=2)$$

$$= (0-1)^2 \bullet \frac{1}{4} \qquad + (1-1)^2 \bullet \frac{1}{2} \qquad + (2-1)^2 \bullet \frac{1}{4}$$

$$= \frac{1}{2}$$

OR

$$E(Y^2) = 0^2 \bullet P(Y=0) + 1^2 \bullet P(Y=1) + 2^2 \bullet P(Y=2)$$

$$= 0 \bullet \frac{1}{4} \quad + \quad 1 \bullet \frac{1}{2} \quad + \quad 4 \bullet \frac{1}{4} \quad = \quad \frac{3}{2}$$

$$Var(Y) = E(Y^2) - [E(Y)]^2 = \frac{3}{2} - 1^2 = \frac{1}{2}$$

Properties of Variance
Var(X±a) = Var(X)
Var(aX) = a^2**Var(X)*

Var(X+Y) = Var(X) + Var(Y) if X & Y are <u>independent</u>

*Var(X+Y) = Var(X) + Var(Y) + 2*Cov(X,Y)* always

*Cov(X,Y) = E([X–E(X)] * [Y–E(Y)])*
*= E(XY) – E(X)*E(Y)*

$$Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} N(\mu, \sigma^2) \qquad \rightarrow \quad \text{a } \textit{Simple Random Sample} \text{ (SRS) of size } n$$

i.i.d. → independent and identically distributed

$$E(\bar{Y}) = E\left(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right) = \frac{1}{n} E\left(Y_1 + Y_2 + \ldots + Y_n\right) = \frac{1}{n}(n \bullet \mu) = \mu$$

$$Var(\bar{Y}) = Var\left(\frac{Y_1 + Y_2 + \ldots + Y_n}{n}\right) = \frac{1}{n^2} Var\left(Y_1 + Y_2 + \ldots + Y_n\right) = \frac{1}{n^2}\left(n \bullet Var(Y_i)\right) = \frac{\sigma^2}{n}$$

$$\frac{\bar{Y} - \mu}{\sigma / \sqrt{n}} = Z \sim N(0,1)$$

2-Sample Tests

$$Y_{11}, Y_{12}, \ldots, Y_{1n_1} \overset{iid}{\sim} N(\mu_1, \sigma_1^2) \qquad H_0 : \mu_1 = \mu_2$$

$$Y_{21}, Y_{22}, \ldots, Y_{2n_2} \overset{iid}{\sim} N(\mu_2, \sigma_2^2)$$

$$E(\bar{Y}_1 - \bar{Y}_2) = E(\bar{Y}_1) - E(\bar{Y}_2) = \mu_1 - \mu_2$$

$$Var(\bar{Y}_1 - \bar{Y}_2) = Var(\bar{Y}_1) + Var(\bar{Y}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

H_o: μ=165 *mg/dL*

H_a: μ>165

Sample Results: A SRS of $n = 30$ gives $\overline{X} = 180.52$

$\sigma = 41.23$ (assume this is known)

a) Find the sample z-score (z_S). [uses 180.52]

b) Find the p-value.

c) State a conclusion for the test at the $\alpha = .05$ level.

1) Write the rejection rule (RR) for H_o in terms of z-scores. [does NOT use 180.52]

2) Write the rejection rule (RR) for H_o in terms of $\overline{X}$.

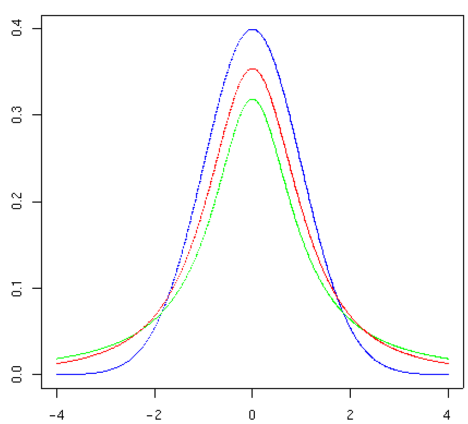
3) Find the probability of a Type II error if \mu=170.□

Four Key Probability Distributions

1. Normal Distribution (assuming σ is a known *population parameter*)
   - CLT → Sampling Distribution for $\bar{X}$ is approximately $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
   - $z_s = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$ has a ***N(0,1)*** distribution

2. Student's T Distribution (assuming σ is unknown)
   - we estimate σ with $s = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$
   - $t_s = \dfrac{\bar{X} - \mu}{s / \sqrt{n}}$ has a *Student's T distribution* with *n-1* degrees of freedom
   - Standard Deviation (of the mean) vs. Standard Error (of the mean)

     $SD_{\bar{X}} = \dfrac{\sigma}{\sqrt{n}}$ $SE_{\bar{X}} = \dfrac{s}{\sqrt{n}}$



Normal Distribution in blue

T Distribution df = 5 in green

T Distribution df = 10 in red

3. Chi-Square Distribution $\left(\chi^2\right)$

- Variances follow a scaled Chi-Square Distribution
- $z_i = \dfrac{Y_i - \mu}{\sigma} \overset{iid}{\sim} N(0,1) \quad \rightarrow \quad z_i^2 = \left(\dfrac{Y_i - \mu}{\sigma}\right)^2 \sim \chi_1^2$
- $z_1^2 + z_2^2 \sim \chi_2^2 \quad \rightarrow \quad \sum_{i=1}^n z_i^2 \sim \chi_n^2 \quad \rightarrow \quad \sum_{i=1}^n \dfrac{(Y_i - \mu)^2}{\sigma^2} \sim \chi_n^2$
- Estimating μ with $\hat{\mu} = \bar{Y} \quad \rightarrow \quad \sum_{i=1}^n \dfrac{(Y_i - \bar{Y})^2}{\sigma^2} \sim \chi_{n-1}^2$

$$S^2 = \frac{\sum (Y_i - \bar{Y})^2}{n-1} \quad \rightarrow \quad \frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$$

4. F Distribution

- A ratio of two scaled Chi-square random variables
- $\dfrac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2 \quad \rightarrow \quad S_1^2 = \dfrac{\sigma_1^2 \chi_{n_1-1}^2}{n_1 - 1} \qquad S_2^2 = \dfrac{\sigma_2^2 \chi_{n_2-1}^2}{n_2 - 1}$
- $\dfrac{S_1^2}{S_2^2} = \dfrac{\sigma_1^2 \chi_{n_1-1}^2 / (n_1 - 1)}{\sigma_2^2 \chi_{n_2-1}^2 / (n_2 - 1)} \overset{*}{=} \dfrac{\chi_{n_1-1}^2 / (n_1 - 1)}{\chi_{n_2-1}^2 / (n_2 - 1)} = F_{n_1-1, n_2-1} \qquad * under\ H_o : \sigma_1^2 = \sigma_2^2$
- If T is *Student's T* random variable with *m* df, then $T^2 \sim F_{1,m}$