STA261 LECTURES NOTES, SPRING 2004

Jeffrey Rosenthal, University of Toronto

(Last updated: March 31, 2004.)

Note: I have decided to make these lecture notes available for STA261 students, for their convenience. I will update them regularly. However, they are just rough, point-form notes, with no guarantee of completeness or accuracy. They should in no way be regarded as a substitute for attending the lectures and tutorials, or for doing the weekly homework exercises.

- Introduction to course, handout, web page, etc.
- How many in Statistics Specialist program? Statistics Major? Actuarial Science? Math? Computer Science? Physics/Chemistry? Economics? Management? Life Sciences? Engineering? Other?
- IDEA OF STATISTICAL INFERENCE: Drawing inference about unknown quantities in the presence of randomness. Uses lots of probability theory!
- INFERENCE WHEN PROBABILITY DISTRIBUTION IS KNOWN (Sect. 5.2):
  - Example (text): X = machine's lifetime in years. Suppose $X \sim \text{Exp}(1)$. This means $P(X > x) = e^{-x}$ for $x \geq 0$. Then $P(X > 5) = e^{-5} \approx 0.0067$. Small! So, machine usually won't last five years. But $P(X > 2) = e^{-2} \approx 0.1353$, not so small. ["Machine lasting 2 years is feasible, lasting 5 years is infeasible"]
  - Example: Suppose patients with disease "Statitus" have 50% chance of dying. [Like flipping coin, with heads=live, tails=die; do experiment.] Then given 8 patients, probability they ALL live is $(1/2)^8 = 1/256 \approx 0.0039$. So, they probably won't all live! But, probability first three live is $(1/2)^3 = 1/8 = 0.125$, not so small. ["First three surviving is feasible, first eight surviving is infeasible"]
  - Example: Roll 6-sided die, patient dies if get 1 or 2 (do experiment). Then probability first two patients die is $(2/6)^2 = 1/9 \approx 0.1111$. Not so unlikely; might happen. But probability first five patients die is $(2/6)^5 = 1/243 \approx 0.0041$, very small. ["First two surviving is feasible, first five surviving is infeasible"]
- INFERENCE WHEN PROBABILITY DISTRIBUTION UNKNOWN (Sect. 5.3):

- Example: Suppose patients with disease "Statitus" are given a new treatment. They then either have 50% chance of dying, or they will all live, but we're not sure which. [Like flipping either regular or two-headed coin; do experiment.] Suppose first 4 patients all live. Does that mean all patients will live? [Probability it happened by chance is $(1/2)^4 = 1/16 = 0.0625$.] "Hypothesis testing".

- Example: Suppose we roll a 6-sided die, patient dies if get one of "certain numbers" (secret). [Do experiment.] What is prob that patient dies? Unknown! Given some observations, how can we ESTIMATE this probability?? "Estimation".

- Example: Suppose you're shooting foul shots in basketball. Your probability p of scoring a basket is unknown. How to estimate it? e.g. Suppose you shoot 10 shots and score 7 times; does that mean $p = 0.7$? Exactly? Are you sure? "Confidence Intervals".

- STATISTICAL MODELS (Sect. 5.3):

  - If probability distribution is unknown, then need to consider various possible probability distributions.

  - Write collection of possible probability distributions as $\{P_\theta : \theta \in \Omega\}$, where θ is a parameter, Ω is the set of possible parameter values, and for each $\theta \in \Omega$, P_θ is a probability distribution on the set S of possible outcomes (or, "responses").

  - For "Statitus treatment" example, could let $S = \{\text{live}, \text{die}\}$, and $\Omega = \{1, 2\}$, and $P_1(\text{die}) = P_1(\text{live}) = 1/2$, and $P_2(\text{live}) = 1$.

  - For "secret list" 6-sided die example, could let $S = \{\text{live}, \text{die}\}$, and $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$, and for $\theta \in \Omega$, $P_\theta(\text{die}) = \theta/6$ and $P_\theta(\text{live}) = 1 - \theta/6$.

  - For basketball example, could let $S = \{\text{score}, \text{miss}\}$, and $\Omega = [0, 1]$, and for $\theta \in \Omega$, $P_\theta(\text{score}) = \theta$ and $P_\theta(\text{miss}) = 1 - \theta$.

- Also need to collect and describe data, with e.g. histograms, etc. (Sect. 5.4 – not emphasised now; maybe later.)

- SOME BASIC METHODS OF INFERENCE (Sect. 5.5.1)

- Suppose we have a random response X whose distribution is unknown. We collect some observations ("data") $x_1, \ldots, x_n$.

- Example: Suppose we're measuring student heights (in centimeters), and we observe: 170, 160, 165, 160, 150, 170.

- Could estimate $F_X(x) = P(X \leq x)$ by $\hat{F}_X(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$, i.e. the *fraction* of observations which are $\leq x$. (Accurate?) [In above example, could estimate that 2/3 of students have height ≤ 165.]

- Could estimate the *mean* ("location parameter") of X by the "sample mean" $\overline{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$. [In above example, average student height is about $\overline{x} = 162.5$.]

—— **END MONDAY 1** ——

[Offer extra handouts as needed.]

[Remind students about suggested homework, posted on website on Thursdays.]

Previous Class:
∗ Inference when probability distribution KNOWN
—— What outcomes are "feasible"?
∗ Introduction to inference when probability distribution UNKNOWN
—— hypothesis testing (e.g. statitus treatment: 50-50 or 100%?)
—— estimation (e.g. # numbers on secret list, when 2/8 die. 2!!)
—— confidence (if you score 7/10 foul shots, how close to 0.7 is p?)
∗ Statistical Models
—— Collection $\{P_\theta : \theta \in \Omega\}$ of possible probability distributions on outcome space S.
—— e.g. $S = \{\text{live, die}\}$, $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$, $P_\theta(\text{die}) = \theta/6$, $P_\theta(\text{live}) = 1 - \theta/6$.

- Some Basic Methods of Inference (Continued)

  - Have a random response X (distribution unknown). Have observations ("data") $x_1, \ldots, x_n$.

  - Example: Measuring student heights (in cm), and observe: 170, 160, 165, 160, 150, 170.

  - Could estimate $F_X(x) = P(X \leq x)$ by $\hat{F}_X(x) \equiv \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,x]}(x_i)$. [In above example, estimate that 2/3 of students have height ≤ 165.]

- Could estimate the *mean* ("location parameter") of X, i.e. $\mu_X = E[X]$, by the "sample mean" $\overline{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$. [In above example, average student height estimated by $\overline{x} = 162.5$.]

- Could estimate the *variance* ("scale") of X, i.e. $\mathrm{Var}(X) = E[(X-\mu_X)^2]$, by the "sample variance" $s^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$. [Why $n-1$ instead of n? Later!] [In above example, variance of student heights estimated by $s^2 = 57.5$.]

- Then estimate *standard deviation* by "sample standard deviation" $s \equiv \sqrt{s^2}$. [In above example, standard deviation estimated by $s = 7.6$.]

- If X is discrete, could estimate $f_X(x) = P(X = x)$ by $\hat{f}_X(x) \equiv \frac{1}{n}\sum_{i=1}^{n} I_x(x_i)$, i.e. the *fraction* of observations which are $= x$. [In above example, should probably *not* conclude that 1/3 of students have height exactly 160, since heights are continuous ...]

- Example: Suppose three candidates (A, B, and C) are running for student president. We select students at random and ask who they will vote for, and observe: A, C, A, B, A, C, A. Then could estimate popularity of candidate A as 4/7, B as 1/7, and C as 2/7. [Here mean, etc. do not make sense, since data are catagorical, i.e. not quantitative.]

- Example: Suppose a random sample of residents are asked to preview a movie and rate it on a scale from 1 to 5. We observe ratings of 4, 2, 1, 3, 2, 1, 4, 2. Then we might estimate that in the general population, 2/8 of people will rate the movie a 1, while 5/8 of people will rate the movie a 1 or 2, and 6/8 of people will rate the movie a 1 or 2 or 3, etc. Also mean rating $\approx \overline{x} = 2.375$, with variance $\approx s^2 \doteq 1.41$, and standard deviation $\approx s \doteq 1.19$. [Movie probably won't be a hit!]

- [Quantile estimation? Omit for now.]

- But how "good" are these estimates??

—— **END WEDNESDAY 1** ——

[Announce tutorial rooms.]

[Reminder re homework to discuss in tutorial: 5.1.1, 5.1.5, 5.1.7, 5.2.4, 5.2.6, 5.2.10, 5.3.1, 5.3.2, 5.3.3, 5.3.5 (model only), 5.5.1 (omit (e)), 5.5.2 (omit (d)).]

[Note: My lecture notes are now on the web page.]

Previous Class:
∗ Examples of basic inference from data:
—— estimate mean by sample mean $\overline{x}$
—— estimate variance by sample variance s^2
—— estimate probabilities and/or cdfs by "fraction of observations"

- LIKELIHOOD FUNCTIONS and MLE (Sect. 6.1, 6.2).
  - Let $\{P_\theta : \theta \in \Omega\}$ be a statistical model on some outcome space S. Suppose we observe some outcome $s \in S$.
- If S is discrete, then the Likelihood Function is the function $L(\cdot \,|\, s)$ on Ω defined by $L(\theta \,|\, s) = P_\theta(s)$, i.e. the probability of observing s if P_θ is the true probability distribution.
  - L is function of parameter θ, given the (fixed) observation s.
  - $L(\theta \,|\, s)$ provides some indication (?) of how "likely" the distribution P_θ is, given the observation s.
  - Example (text): $S = \{1, 2, 3, \ldots\}$, $\Omega = \{1, 2\}$, $P_1 = \text{Uniform}\{1, 2, \ldots, 1000\}$, and $P_2 = \text{Uniform}\{1, 2, \ldots, 1000000\}$. Observe $s = 10$. Then $L(1 \,|\, 10) = 1/1000$, and $L(2 \,|\, 10) = 1/1000000$. Suggests that P_1 much more likely than P_2, even though both values very small.
  - Definition: The Maximum Likelihood Estimator" (MLE) of θ is the value of θ which maximises $L(\theta \,|\, s)$. In above example, MLE is $\hat{\theta} = 1$.
  - Example: "Statitus treatment" example: $S = \{\text{live, die}\}$, $\Omega = \{1, 2\}$, $P_1(\text{live}) = P_1(\text{die}) = 1/2$, $P_2(\text{live}) = 1$. If we observe $s = \text{live}$, then $L(1 \,|\, \text{live}) = 1/2$, $L(2 \,|\, \text{live}) = 1$, so P_2 more likely (in fact, twice!). But if we observe $s = \text{die}$, then $L(1 \,|\, \text{die}) = 1/2$, $L(2 \,|\, \text{live}) = 0$, so P_1 more likely (in fact, infinitely more!). So, if

$s = \text{live}$ then MLE is $\hat{\theta} = 2$, but if $s = \text{die}$ then MLE is $\hat{\theta} = 1$.

- Can also compute likelihood under <u>multiple</u> observations. In "Statitus treatment" example, if observation s corresponds to <u>three</u> patients who <u>all</u> live, then $P_1(s) = (1/2)^3 = 1/8$, while $P_2(s) = 1$. So can write $L(1 \mid s) = 1/8$, $L(2 \mid s) = 1$, so P_2 is <u>eight times</u> more likely, and MLE is $\hat{\theta} = 2$.

- e.g. "secret list" 6-sided die example, where $S = \{\text{live, die}\}$, $\Omega = \{0, 1, 2, 3, 4, 5, 6\}$, and $P_\theta(\text{die}) = \theta/6$. If observe one patient die, then $L(\theta \mid \text{die}) = \theta/6$ for $\theta \in \Omega$, largest at $\hat{\theta} = 6$. If observe one patient live, then $L(\theta \mid \text{live}) = 1 - \theta/6$ for $\theta \in \Omega$, largest at $\hat{\theta} = 0$.

- If instead observation s is that 2 out of 8 patients died, then $L(\theta \mid s) = \binom{8}{2}(\theta/6)^2(1 - \theta/6)^6 = 28(\theta/6)^2(1 - \theta/6)^6$. Thus, $L(0 \mid s) = 0$, $L(1 \mid s) = 28(1/6)^2(1 - 1/6)^6 \doteq 0.260$, $L(2 \mid s) \doteq 0.273$, $L(3 \mid s) \doteq 0.109$, $L(4 \mid s) \doteq 0.017$, $L(5 \mid s) \doteq 0.0004$, $L(6 \mid s) \doteq 0$. Suggests $\theta = 2$ is most likely (was actually true!), so MLE is $\hat{\theta} = 2$, though $\theta = 1$ fairly likely too. ($\theta = 3$ less so.)

- Comment: Two different likelihood functions are <u>equivalent</u> (i.e., just as good) if one is a positive constant times the other [since we only care about the <u>ratios</u> $L(\theta_1 \mid s) \,/\, L(\theta_2 \mid s)$]. So, in above example, could have ignored the "28" if we wanted. More generally, can ignore <u>any</u> positive factor which does not depend on θ (even if it depends on the observation s).

• If S is <u>continuous</u>, so each P_θ has a <u>density</u> f_θ, then can define likelihood function by $L(\theta \mid s) = f_\theta(s) =$ value of density function. (Note: In discrete case, sometimes also write $f_\theta(s)$ for $p_\theta(s)$, i.e. for $P_\theta[s]$.)

- Example ("one Normal observation"): Suppose $S = \mathbf{R}$, and $\Omega = \mathbf{R}$, and $P_\theta = N(\theta, 1) =$ normal distribution. Thus $f_\theta(s) = \frac{1}{\sqrt{2\pi}} e^{-(s-\theta)^2/2}$. If we observe $s \in S$, then $L(\theta \mid s) = \frac{1}{\sqrt{2\pi}} e^{-(s-\theta)^2/2}$. Equivalently, can take $L(\theta \mid s) = e^{-(s-\theta)^2/2}$. Largest when $\theta = s$, so MLE is $\hat{\theta} = s$. (Makes sense ...)

- Example ("one Exponential observation"): Suppose $S = (0, \infty)$ and $\Omega = (0, \infty)$, with $P_\theta = \text{Exp}(\theta)$, and we observe one outcome $s > 0$. Then $L(\theta \mid s) = f_\theta(s) = \theta e^{-s\theta}$. How to maximise?

– Well,

$$\frac{\partial}{\partial\theta}L(\theta\,|\,s) = e^{-s/\theta} - \theta e^{-s/\theta}(s)$$

which equals 0 iff $1 - s\theta = 0$, i.e. $\theta = 1/s$. This appears to maximise $L(\theta\,|\,s)$, so that MLE is $\hat{\theta} = s$. (Makes sense since mean of $\mathrm{Exp}(\theta)$ is $1/\theta$, so mean of $\mathrm{Exp}(1/s)$ is s ...)

– <u>Easier</u> is to consider <u>logarithm</u> of likelihood; since logarithm is an <u>increasing</u> function, maximising log-likelihood is same as maximising likelihood. Compute:

$$\ell(\theta\,|\,s) = \log[L(\theta\,|\,s)] = \log[\theta e^{-s/\theta}] = \log(\theta) - s\theta\,.$$

Then derivative of this is the <u>score function</u>:

$$S(\theta\,|\,s) = \frac{\partial}{\partial\theta}\ell(\theta\,|\,s) = \frac{\partial}{\partial\theta}[\log(\theta) - s\theta] = 1/\theta - s\,,$$

and this equals 0 ["Score Equation"] if and only if $(1/\theta) - s = 0$, i.e. $\theta = 1/s$.

– As a check, the <u>second</u> derivative is $\frac{\partial}{\partial\theta}\ell(\theta\,|\,s) = -\theta^{-2}$. At $\theta = \hat{\theta} = s$, this equals $s^{-2} < 0$. Hence, $\theta = \hat{\theta}$ is indeed a <u>local maximum</u>, and then easily seen to be a global maximum.

– Example ("multiple Exponential observations"): Again $S = (0,\infty)$ and $\Omega = (0,\infty)$, with $P_\theta = \mathrm{Exp}(\theta)$, and we observe n outcomes $x_1, x_2, \ldots, x_n > 0$. Then

$$L(\theta\,|\,x_1,\ldots,x_n) = \prod_{i=1}^{n}[\theta e^{-x_i\theta}] = \theta^n e^{-\sum_{i=1}^{n} x_i\theta} = \theta^n e^{-n\overline{x}\theta}\,.$$

Then

$$\ell(\theta\,|\,x_1,\ldots,x_n) = \log[L(\theta\,|\,s)] = n\log(\theta) - n\overline{x}\theta\,.$$

Hence, score function is

$$S(\theta\,|\,x_1,\ldots,x_n) = \frac{\partial}{\partial\theta}\ell(\theta\,|\,x_1,\ldots,x_n) = n/\theta - n\overline{x}\,,$$

which equals 0 iff $(1/\theta) - \overline{x} = 0$, i.e. $\theta = 1\,/\,\overline{x}$. (Makes sense, since could also estimate mean $1/\theta$ by $\overline{x}$, equivalent to estimating θ by $1/\overline{x}$.)

—— **END MONDAY 2** ——

[Reminder about tutorials today, after lecture.]

Previous Class:
* Likelihood function $L(\theta \,|\, s)$.
—— Indicates relative likelihood of P_θ being true, given observation s.
* Discrete case: $L(\theta \,|\, s) = P_\theta(s)$ (probability). [Examples.]
* (Absolutely) continuous case: $L(\theta \,|\, s) = f_\theta(s)$ (density). [Examples.]
* MLE is value of θ which <u>maximises</u> $L(\theta \,|\, s)$.
* Two likelihood functions are <u>equivalent</u> if $L_1(\theta \,|\, s) = K L_2(\theta \,|\, s)$ for all $\theta \in \Omega$, for some $K > 0$ which does not depend on θ.

- Aside about likelihood equivalence: L_1 and L_2 are equivalent iff the ratio L_1/L_2 does not depend on θ. For example, suppose $L_1(\theta \,|\, s) = \theta^2$, $L_2(\theta \,|\, s) = 15\theta^2$, $L_3(\theta \,|\, s) = s^3\theta^2$, $L_4(\theta \,|\, s) = \theta$. Which are equivalent? Answer: L_1, L_2, and L_3 are equivalent, but L_4 is <u>not</u>. So, can't just erase a constant (like 2) from the <u>exponent</u>. Similarly, if $L_5(\theta \,|\, s) = e^{-\theta}$ and $L_6(\theta \,|\, s) = e^{-\theta/2}$, then $L_5(\theta \,|\, s)/L_6(\theta \,|\, s) = e^{-\theta/2}$, which depends on θ, so L_5 and L_6 <u>not</u> equivalent.

- Likelihood functions, continuous case (continued).

  – Example ("multiple Normal observations"): Suppose observe <u>multiple</u> data $x_1, x_2, \ldots, x_n$ from $N(\theta, 1)$. Then can take $L(\theta \,|\, x_1, \ldots, x_n) = \prod_{i=1}^n e^{-(x_i-\theta)^2/2} = \exp\big(-\frac{1}{2}\sum_{i=1}^n (x_i - \theta)^2\big)$.

  – In fact, above likelihood function is <u>equivalent</u> to $L_2(\theta \,|\, x_1, \ldots, x_n) = \exp\big(-\frac{n}{2}(\overline{x} - \theta)^2\big)$. Proof:

$$\sum_{i=1}^n (x_i - \theta)^2 = \sum_{i=1}^n (x_i^2 - 2x_i\theta + \theta^2) = (\sum_{i=1}^n x_i^2) - 2n\overline{x}\theta + n\theta^2$$

  while

$$n(\overline{x} - \theta)^2 = n\overline{x}^2 - 2n\overline{x}\theta + n\theta^2$$

  so difference between them is

$$(\sum_{i=1}^n x_i^2) - n\overline{x}^2$$

  which does not depend on θ. So,

$$\frac{L(\theta \,|\, s)}{L_2(\theta \,|\, s)} \;=\; \exp\big(-\frac{1}{2}((\sum_{i=1}^n x_i^2) - n\overline{x}^2)\big)\,,$$

which does not depend on n. ■

- Hence, MLE is $\hat{\theta} = \overline{x}$.

- "Uniform" Example (text): Suppose model is $S = [0, \infty)$, $\Omega = (0, \infty)$, and $P_\theta = \text{Uniform}[0, \theta]$, and we observe $x_1, x_2, \ldots, x_n \geq 0$. How to estimate θ? Here $L(\theta \,|\, x_1, \ldots, x_n) = 1/\theta^n$ <u>if</u> $0 \leq x_i \leq \theta$ for all i, otherwise $L(\theta \,|\, x_1, \ldots, x_n) = 0$. By observation (not differentiation!), this is maximised at $\theta = \hat{\theta} = \max\{x_i;\ 1 \leq i \leq n\}$. This is the MLE.

- If instead $S = (-\infty, \infty)$, $\Omega = (0, \infty)$, and $P_\theta = \text{Uniform}[-\theta, \theta]$, and observe $x_1, x_2, \ldots, x_n$, then MLE is $\hat{\theta} = \max\{|x_i|;\ 1 \leq i \leq n\}$. (exercise)

- Example ("Multinomial Model"): Suppose individual responses can take one of the values $S = \{1, 2, \ldots, k\}$ (e.g. election preference; perhaps $k = 3$), with various probabilities (unknown). ["Catagorical response".] Statistical model is

$$\Omega = \{\theta = (\theta_1, \theta_2, \ldots, \theta_k)\,;\ \theta_i \geq 0,\ \theta_1 + \ldots + \theta_k = 1\}\,,$$

and $P_\theta(i) = \theta_i$. If we observe responses $x_1, x_2, \ldots, x_n$ (perhaps n is large), then likelihood function is

$$L(\theta \,|\, x_1, x_2, \ldots, x_n) \;=\; \theta_{x_1} \theta_{x_2} \ldots \theta_{x_n}\,.$$

This is equal to $\theta_1^{c_2} \theta_2^{c_2} \ldots \theta_k^{c_k}$, where $c_i = \#\{j : x_j = i\}$ = <u>count</u> of number of responses of type i. Hence, likelihood <u>only depends</u> on the <u>count</u> data $(c_1, c_2, \ldots, c_k)$, not on the full response list $(x_1, x_2, \ldots, x_n)$.

—— **END WEDNESDAY 2** ——

[Some office hours now posted on web site (TA's, plus New College). Also a few "extra" hours available per TA. However, these office hours are to SUPPLEMENT the tutorials, not REPLACE them!]

Previous Class:
* Example re likelihood equivalence.
* MLE for Multiple Normal observations.
* MLE for Multiple Uniform observations.
* Multinomial Model: $P_\theta(i) = \theta_i$, $L(\theta \,|\, x_1, \ldots, x_n) = \theta_{x_1} \ldots \theta_{x_n} = (\theta_1)^{c_1} \ldots (\theta_k)^{c_k}$

- Numerical example for Multinomial Model: Suppose $k = 3$, and $n = 7$, and observations are $1, 2, 3, 2, 2, 1, 2$. Then

$$L(\theta \,|\, 1, 2, 3, 2, 2, 1) = \theta_1\theta_2\theta_3\theta_2\theta_2\theta_1\theta_2 = (\theta_1)^2(\theta_2)^4(\theta_3)^1 \,,$$

  so here $c_1 = 2$, $c_2 = 4$, and $c_3 = 1$. Can thus summarise the full observation list $(1, 2, 3, 2, 2, 1, 2)$ by the count data $(2, 4, 1)$.

- SUFFICIENT STATISTICS (6.1.1).
  - Definition: A statistic is some function T of the data $(x_1, \ldots, x_n)$, e.g. $\overline{x}$, s^2, $\frac{1}{n}\sum_{i=1}^{n} I_{(\infty,5]}(x_i)$, the count data $(c_1, c_2, \ldots, c_k)$, etc.
  - Definition: A statistic T is sufficient if different observations, with the same value of the statistic, always have equivalent likelihood functions. i.e., if whenever $T(s_1) = T(s_2)$, then $L(\theta \,|\, s_1) = K\, L(\theta \,|\, s_2)$ for all $\theta \in \Omega$, for some constant $K > 0$ (which may depend on s_1 and s_2).
  - In above "Multinomial Model" example, the statistic of "count data", i.e. $T(x_1, x_2, \ldots, x_n) = (c_1, c_2, \ldots, c_k)$, is sufficient, since the likelihood function only depends on $(c_1, c_2, \ldots, c_k)$.
  - In "Normal Observations" example, likelihood function is equivalent to $L(\theta \,|\, x_1, \ldots, x_n) = \exp\big(-\frac{n}{2}(\overline{x} - \theta)^2\big)$, so L only depends on the data through $\overline{x}$, hence the statistic $\overline{x}$ is sufficient.
  - "a-b Example" (text): Let $S = \{1, 2, 3, 4\}$, and $\Omega = \{a, b\}$, with $P_a(1) = 1/2$ and $P_a(2) = P_a(3) = P_a(4) = 1/6$, and with $P_b(1) = P_b(2) = P_b(3) = P_b(4) = 1/4$. What is a sufficient statistic? Well, note that $L(\theta \,|\, s) = P_\theta(s)$ is the same if s equals 2, 3, or 4. Hence, likelihood "does not care" if observation is 2, 3, or 4. So, let $T : S \to \{0, 1\}$ by $T(1) = 0$, and $T(2) = T(3) = T(4) = 1$. Then if $T(s_1) = T(s_2)$, then $L(\theta \,|\, s_1) = L(\theta \,|\, s_2)$. Hence, T is sufficient statistic.

- FACTORISATION THEOREM: Let $f_\theta(s)$ be probability (or density function) for a statistical model, and let T be a statistic. Suppose can "factor" $f_\theta(s)$ as $f_\theta(s) = h(s)\, g_\theta(T(s))$ for some positive functions g_θ and h. [Often take $h(s) \equiv 1$.] Then T is a sufficient statistic.

- Proof: If $T(s_1) = T(s_2)$, then

$$L(\theta \,|\, s_1) = f_\theta(s_1) = h(s_1) g_\theta(T(s_1)) = h(s_1) g_\theta(T(s_2))$$

$$= \frac{h(s_1)}{h(s_2)} h(s_2) g_\theta(T(s_2)) = \frac{h(s_1)}{h(s_2)} L(\theta \,|\, s_2) = K\, L(\theta \,|\, s_2)\,,$$

where $K = h(s_1)/h(s_2)$ does not depend on θ. ∎

- In above "a-b Example", can write $f_\theta(s) = 1 \cdot g_\theta(T(s))$ where $g_a(0) = 1/2$, $g_a(1) = 1/6$, $g_b(0) = g_b(1) = 1/4$. So T is sufficient statistic.

• A statistic T is a minimal sufficient statistic if $T(s_1) = T(s_2)$ if and only if $L(\theta \,|\, s_1) = L(\theta \,|\, s_2)$ $\forall \theta \in \Omega$, i.e. we can calculate $T(s)$ once we know the mapping $\theta \mapsto L(\theta \,|\, s)$.

- Intuitively, this means T is a "best possible" sufficient statistic.

- In above "a-b Example", $L(a \,|\, s) = 1/2$ if $T(s) = 0$, while $L(a \,|\, s) = 1/6$ if $T(s) = 1$, so T is minimal sufficient statistic.

- Similarly, for Multinomial Model, $(c_1, \ldots, c_k)$ is minimal sufficient statistic; and for Normal Observations example, $\overline{x}$ is minimal sufficient statistic. (Exercise.)

• REPARAMETERIZATION (6.2): Given statistical model $\{P_\theta : \theta \in \Omega\}$, suppose $\Psi : \Omega \to \Omega'$ is 1–1, Then MLE of new parameter $\psi \equiv \Psi(\theta)$ is given by $\hat{\psi} \equiv \Psi(\hat{\theta}(s))$. ["Plug-in estimator"]

- Multiple Uniform Example: $S = [0, \infty)$, $\Omega = (0, \infty)$, $P_\theta = \text{Uniform}[0, \theta]$, observe $x_1, \ldots, x_n$. We know MLE of θ is $\hat{\theta} = \max_{1 \le i \le n}\{x_i\}$. Thus, since $\theta \mapsto e^\theta$ is 1–1, MLE of e^θ is $\widehat{e^\theta} = e^{\hat{\theta}} = \exp\big(\max_{1 \le i \le n}\{x_i\}\big) = \max_{1 \le i \le n}\{e^{x_i}\}$. Also, $\theta \mapsto \theta^2$ is 1-1 on Ω, so MLE of θ^2 is $\widehat{\theta^2} = (\hat{\theta})^2 = \max_{1 \le i \le n}\{(x_i)^2\}$. However, MLE of $(\theta - 5)^2$ is unclear since function is not 1–1.

• ESTIMATOR BIAS (6.3.1):

- Given estimator $\hat{\theta}$ of θ, how good is it?

- Write $E_\theta(\hat{\theta})$ for the expected value of $\hat{\theta}$, under the distribution P_θ, i.e. assuming that θ is the true parameter value.

- The bias of the estimator is $\text{Bias}_\theta(\hat{\theta}) = E_\theta(\hat{\theta}) - \theta$.

– Example: Suppose $S = [0, 1]$, $\Omega = \{1, 2\}$, and $f_1(s) = 1$ and $f_2(s) = 2s$ for $s \in S$. The MLE of θ is $\hat{\theta} = 1$ if $s < 1/2$, while $\hat{\theta} = 2$ if $s > 1/2$. [If $s = 1/2$, MLE is *either* 1 or 2.] Now, $P_1[s < 1/2] = P_2[s > 1/2] = 1/2$, so $E_1(\hat{\theta}) = 3/2$, so $\text{Bias}_1(\hat{\theta}) = (3/2) - 1 = +1/2$. Also $P_2[s < 1/2] = \int_0^{1/2} 2s\, ds = (1/2)^2 = 1/4$ and $P_2[s > 1/2] = \int_{1/2}^1 2s\, ds = 3/4$, so $E_2(\hat{\theta}) = (1/4)(1) + (3/4)(2) = 7/4$, and $\text{Bias}_2(\hat{\theta}) = (7/4) - 0.5 = -1/4$.

—— **END MONDAY 3** ——

[Kung Hay Fat Choy!]

Previous Class:
* Sufficient Statistics
* Factorisation Theorem
* Minimal Sufficient Statistics
* Reparameterisation
* Estimator Bias

- Another example re Factorisation Theorem & Minimal Sufficient Statistics: $S = \Omega = \mathbf{R}$, $P_\theta = N(\theta, 1)$, and $L(\theta \,|\, x_1, \ldots, x_n) = \exp(-(n/2)(\overline{x} - \theta)^2)$.

  – Then $L(\theta \,|\, x_1, \ldots, x_n) = h(x_1, \ldots, x_n)\, g_\theta(T(x_1, \ldots, x_n))$, where $T(x_1, \ldots, x_n) = \frac{1}{n}(x_1 + \ldots + x_n) = \overline{x}$, $h(x_1, \ldots, x_n) \equiv 1$, and $g_\theta(r) = \exp(-(n/2)(r - \theta)^2)$. Hence, by Factorisation Theorem, $\overline{x}$ is sufficient statistic.

  – Is $\overline{x}$ minimal?

  – Suppose have two sets of observations, $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$. Suppose that $L(\theta \,|\, x_1, \ldots, x_n) \propto L(\theta \,|\, y_1, \ldots, y_n)$, i.e. $\exp(-(n/2)(\overline{x} - \theta)^2) = K \exp(-(n/2)(\overline{y} - \theta)^2)$, $\forall \theta \in \Omega$, some $K > 0$. Does this mean that $\overline{x} = \overline{y}$, i.e. $T(x_1, \ldots, x_n) = T(y_1, \ldots, y_n)$?

  – Yes! Theorem: $\overline{x}$ is a <u>minimal</u> sufficient statistic.

  – Proof #1 ("constructive"): If $\exp(-(n/2)(\overline{x} - \theta)^2) = K \exp(-(n/2)(\overline{y} - \theta)^2)$, $\forall \theta \in \Omega$, for some $K > 0$, then both functions must take their <u>maximum</u> at the same value of θ. But LHS takes maximum at $\overline{x}$, while RHS takes maximum at $\overline{y}$. So, must have $\overline{x} = \overline{y}$.

- Proof #2 (by "contraposition", a form of contradiction): Suppose theorem is false. That means we sometimes have $\overline{x} \neq \overline{y}$, even though $\exp(-(n/2)(\overline{x}-\theta)^2) = K \exp(-(n/2)(\overline{y}-\theta)^2)$, $\forall \theta \in \Omega$. Is this possible?? If so, then setting $\theta = \overline{x}$ gives $1 = K \exp(-(n/2)(\overline{y}-\overline{x})^2) < K$, i.e. $K > 1$. But setting $\theta = \overline{y}$ gives $\exp(-(n/2)(\overline{x}-\overline{y})^2) = K$, i.e. $K < 1$. Contradiction! i.e., if $\overline{x} \neq \overline{y}$, then we <u>cannot</u> have $\exp(-(n/2)(\overline{x}-\theta)^2) = K \exp(-(n/2)(\overline{y}-\theta)^2)$, $\forall \theta \in \Omega$. So, theorem must be true, i.e. $\overline{x}$ must be a minimal sufficient statistic.

- Aside re logic: The principle of "contraposition" states: "P implies Q" is equivalent to "not-Q implies not-P"; indeed, both mean it is impossible to have both P true and Q false, at the same time. [Example: "$x > 5$ implies $x > 4$" is equivalent to "$x \leq 4$ implies $x \leq 5$". But <u>not</u> equivalent to "$x > 4$ implies $x > 5$".]

- Note that for $\theta = (\overline{x}+\overline{y})/2$, we do have $\exp(-(n/2)(\overline{x}-\theta)^2) = \exp(-(n/2)(\overline{y}-\theta)^2)$. But not true for <u>all</u> $\theta \in \Omega$.

- Similarly, $x_1 + \ldots + x_n$ is also minimal sufficient statistic, but just x_1 is <u>not</u> sufficient.

- By contrast, if we consider the pair $w = (x_1,\ x_2 + \ldots + x_n)$, then w is still sufficient (since can compute $\overline{x}$ from it), but w is not minimal (since from the likelihood function there is no way to compute x_1, just $\overline{x}$ or $x_1 + \ldots + x_n$).

• Estimator Bias, continued:

- More generally, any parameter $\psi = \Psi(\theta)$ with estimator $\hat{\psi}$ has bias given by $\text{Bias}_\theta(\hat{\psi}) = E_\theta(\hat{\psi}) - \Psi(\theta)$.

- Example: Suppose P_θ has mean $\psi = \Psi(\theta)$ [or just θ], and we estimate ψ by $\overline{x}$. Then $E_\theta(X_i) = \psi$, so $E_\theta(\overline{X}) = \psi$, so $\text{Bias}_\theta(\hat{\psi}) = 0$ no matter what θ is. ["Unbiased Estimator"]

- Multiple Normal Example: Suppose $\Omega = \mathbf{R}$, and $P_\theta = N(\theta, 1)$. Then MLE of θ is $\hat{\theta} = \overline{x}$. Then $E_\theta(\overline{X}) = \theta$, so $\text{Bias}_\theta(\hat{\theta}) = 0$ for all $\theta \in \Omega$. [Good.]

- Multiple Uniform Example: Here $P_\theta = \text{Uniform}[0, \theta]$, and MLE is $\hat{\theta} = \max_{1 \leq i \leq n}\{x_i\}$. Is it unbiased? No, since $P_\theta[\hat{\theta} < \theta] = 1$, so $E_\theta[\hat{\theta}] < \theta$, so $\text{Bias}_\theta(\hat{\theta}) < 0$. (How much less?) ["Biased Estimator"] (Bad??)

—— **END WEDNESDAY 3** ——

[New info available on web.]

Previous Class:

∗ Detailed example of sufficient statistics, factorisation theorem, minimal sufficiency.

∗ More examples about Estimator Bias.

- YET MORE ABOUT BIAS:

- Multiple Uniform Example (cont'd): $P_\theta = \text{Uniform}[0, \theta]$, $\hat{\theta} = \max_{1 \le i \le n}\{x_i\}$. Then $P_\theta[\hat{\theta} < \theta] = 1$, so $E_\theta[\hat{\theta}] < \theta$, so $\text{Bias}_\theta(\hat{\theta}) < 0$. (How much less?) ["Biased Estimator"] (Bad??)

  – Alternate estimator: $\hat{\theta}_2 = 2\overline{x}$. Then $E_\theta(\hat{\theta}_2) = 2\,E_\theta(\overline{x}) = 2(\theta/2) = \theta$, so <u>unbiased</u>. (Good??) But could have $\hat{\theta}_2 < x_i$ for some i. (Crazy??)

- "Location-Scale Normal Model": Suppose $\Omega = \mathbf{R} \times (0, \infty)$, where for $\theta = (\mu, \sigma^2) \in \Omega$, we have $P_\theta = P_{(\mu,\sigma^2)} = N(\mu, \sigma^2)$. i.e. both μ and σ^2 unknown.

  – FACT (Text Example 6.2.6): Here MLE of (μ, σ^2) is $\left(\overline{x},\ \frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2\right)$. [Requires solving a two-dimensional Score Equation.]

  – Thus, $\hat{\mu} = \overline{x}$, which is unbiased.

  – What about estimator of σ^2? Fact (Text Corollary 4.6.2): $E_\theta[\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2] = \frac{n-1}{n}\,\sigma^2$. Thus, always have $E_\theta[\frac{1}{n}\sum_{i=1}^n (x_i - \overline{x})^2] < \sigma^2$ – biased! [Bad??]

  – If instead use $S^2 = \frac{1}{n-1}\sum_{i=1}^n (x_i - \overline{x})^2$, then $E_\theta[S^2] = \frac{n}{n-1}\,\frac{n-1}{n}\,\sigma^2 = \sigma^2$. Thus, S^2 is unbiased estimator of σ^2. [This is why, in S^2, we divide by $n-1$ instead of n.] [In fact, $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, and S^2 independent of $\overline{x}\ldots$]

- MEAN SQUARED ERROR (6.3.1):

- Defn: Let Ψ be a function of a parameter θ, with estimator $\hat{\psi}$. The <u>mean squared error</u> of $\hat{\psi}$ is

$$MSE_\theta(\hat{\psi}) \;=\; E_\theta[(\hat{\psi} - \Psi(\theta))^2]\,, \qquad \theta \in \Omega\,.$$

  (Best if small!)

- Theorem (text Thm 6.3.1): $MSE_\theta(\hat{\psi}) = \mathrm{Var}_\theta(\hat{\psi}) + (\mathrm{Bias}_\theta(\hat{\psi}))^2$.
  - Proof:

$$E_\theta\left(\left(\hat{\psi} - \psi(\theta)\right)^2\right) = E_\theta\left(\left(\hat{\psi} - E_\theta\left(\hat{\psi}\right) + E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right)^2\right)$$

$$= E_\theta\left(\left(\hat{\psi} - E_\theta\left(\hat{\psi}\right)\right)^2\right) + 2\,E_\theta\left(\left(\hat{\psi} - E_\theta\left(\hat{\psi}\right)\right)\left(E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right)\right) + \left(E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right)^2$$

$$= \mathrm{Var}_\theta\left(\hat{\psi}\right) + 2(0) + \left(\mathrm{Bias}_\theta(\hat{\psi})\right)^2,$$

    since

$$E_\theta\left(\left(\hat{\psi} - E_\theta\left(\hat{\psi}\right)\right)\left(E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right)\right) = \left(E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right) E_\theta\left(\left(\hat{\psi} - E_\theta\left(\hat{\psi}\right)\right)\right)$$

$$= \left(E_\theta\left(\hat{\psi}\right) - \psi(\theta)\right)\ (0) = 0\,. \ \blacksquare$$

- Example: $P_\theta = N(\theta, 1)$. Then MLE is $\hat{\theta} = \overline{x}$. Know $\mathrm{Bias}_\theta(\overline{x}) = 0$. Hence $MSE_\theta(\overline{x}) = \mathrm{Var}_\theta(\overline{x}) = 1/n$. [Gets smaller as $n \to \infty$.]

- Example: Conducting referendum. $S = \{\text{yes}, \text{no}\}$. $\Omega = [0, 1]$. $P_\theta(\text{yes}) = \theta$, $P_\theta(\text{no}) = 1 - \theta$. [Just like basketball example.] Observe $x_1, \ldots, x_n$. What is MLE? What is MSE of MLE?
  - Likelihood is $L(\theta \,|\, x_1, \ldots, x_n) = \theta^c (1-\theta)^{n-c}$, where $c = \#\{i;\ x_i = \text{yes}\}$.
  - Then $\ell(\theta \,|\, x_1, \ldots, x_n) = c \log(\theta) + (n-c)\log(1-\theta)$.
  - Then $S(\theta \,|\, x_1, \ldots, x_n) = (c/\theta) - ((n-c)/(1-\theta))$.
  - Score Equation solved when $c(1-\theta) - (n-c)\theta = 0$, i.e. $c - n\theta = 0$, i.e. $\theta = c/n$. So, MLE is $\hat{\theta} = c/n$. [Makes sense.]
  - Also, under P_θ, $c \sim \mathrm{Binomial}(n,\, \theta)$, so $E_\theta(c) = n\theta$, so $E_\theta(c/n) = \theta$, so $\hat{\theta}$ unbiased.
  - Hence, $MSE_\theta(\hat{\theta}) = \mathrm{Var}_\theta(\hat{\theta}) = n\theta(1-\theta)/n^2 = \theta(1-\theta)/n$.
  - Problem: θ unknown!! What to do?
  - Option #1: Note that always have $\theta(1-\theta) \le 1/4$, so must have $MSE_\theta(\hat{\theta}) \le (1/4)/n = 1/4n$. [Conservative estimate; what most polling companies do!]

- Option #2: Instead use the estimated mean squared error $MSE_{\hat{\theta}}(\hat{\theta})$, i.e. $MSE_{\hat{\theta}}(\hat{\theta}) = \hat{\theta}(1-\hat{\theta})/n = (c/n)(1-c/n)/n = c(1-c/n)/n^2$. [Less conservative.]
- Corresponding standard error is then $Sd_{\hat{\theta}}(\hat{\theta}) = \sqrt{MSE_{\hat{\theta}}(\hat{\theta})} = \sqrt{c(1-c/n)/n^2} = \sqrt{c(1-c/n)}\,/\,n$.

- Aside: Predicting weather. Suppose Environment Canada says, "20% chance of rain tomorrow", and then it rains. Are they wrong? How to judge??
  - Using idea of MSE, their "error" equals $(80\%)^2$, i.e. 0.64 error.
  - More generally, if they predict probability p of precipitation (POP), then if it rains or snows their "error" is $(1-p)^2$, otherwise their "error" is p^2. ["Brier Score" ...]
  - Without the square, error is minimised by always predicting either 0% or 100% POP. But with square, error is minimised by best estimate $\hat{p}$ of true probability.

- Example: Suppose $P_\theta = \text{Uniform}[0,\theta]$, and $\hat{\theta} = \max_{1\le i\le n} x_i$. What is $MSE_\theta(\hat{\theta})$?
  - Well, $P_\theta[(\hat{\theta}-\theta)^2 \ge r] = P_\theta[\hat{\theta} \le \theta - \sqrt{r}] = ((\theta-\sqrt{r})/\theta)^n$. So, use trick:

$$MSE_\theta(\hat{\theta}) = E_\theta[(\hat{\theta}-\theta)^2] = \int_0^\theta P_\theta[(\hat{\theta}-\theta)^2 \ge r]\,dr = \int_0^\theta ((\theta-\sqrt{r})/\theta)^n\,dr\,.$$

    [Messy to compute, use computer ...]
  - Suppose $\hat{\theta}_2 = 2\,\overline{x}$. Then $\text{Bias}_\theta(\hat{\theta}_2) = 0$, while $\text{Var}_\theta(\hat{\theta}_2) = (4/n)\text{Var}_\theta(x_i) = \theta^2/3n$, so $MSE_\theta(\hat{\theta}_2) = 0^0 + \theta^2/3n = \theta^2/3n$.
  - e.g. $\theta = 5$, $n = 10$: $MSE_\theta(\hat{\theta}) \doteq 0.38$, $MSE_\theta(\hat{\theta}_2) \doteq 0.84$.
  - e.g. $\theta = 5$, $n = 100$: $MSE_\theta(\hat{\theta}) \doteq 0.005$, $MSE_\theta(\hat{\theta}_2) \doteq 0.084$. ["$\hat{\theta}$ better"?]

- CONSISTENCY: Say an estimator $\hat{\theta}$ of a parameter θ is consistent if, as the number of observations n goes to infinity, $\hat{\theta}$ converges to θ in probability, i.e. for all $\epsilon > 0$, $\lim_{n\to\infty} P_\theta[|\hat{\theta}-\theta| \ge \epsilon] = 0$. [Good.]

- Example: $P_\theta = \text{Uniform}[0,\theta]$, $\hat{\theta} = \max_{1\le i\le n} x_i$, $\hat{\theta}_2 = 2\,\overline{x}$. Are they consistent?
  - By WLLN, as $n\to\infty$, $\overline{x} \to \theta/2$ (mean) in probability. So, $\hat{\theta}_2 \to \theta$ in probability. Consistent!

- What about $\hat{\theta}$? Well, given $\epsilon > 0$, $P_\theta[|\hat{\theta}-\theta| \geq \epsilon] = P_\theta[\hat{\theta} \leq \theta-\epsilon] = ((\theta-\epsilon)/\theta)^n \to 0$ as $n \to 0$. So, $\hat{\theta}$ also consistent.

—— **END MONDAY 4** ——

Previous Class:
$*$ More about bias, S^2.
$*$ Mean Squared Error: $MSE_\theta(\hat{\theta}) = \text{Var}_\theta(\hat{\theta}) + (\text{Bias}_\theta(\hat{\theta}))^2$.
—— Examples: Normal, Referendum, Weather, Uniform
$*$ Consistency: $\hat{\theta} \to \theta$ in probability, as $n \to \infty$.
—— Uniform: Both $\hat{\theta}$ and $\hat{\theta}_2$ consistent.

- Theorem: If $\lim_{n\to\infty} MSE_\theta(\hat{\theta}) = 0$, then $\hat{\theta}$ is a consistent estimator for θ.

  - Proof: By Markov's inequality,

$$P_\theta[|\hat{\theta}-\theta| \geq \epsilon] \;=\; P_\theta[(\hat{\theta}-\theta)^2 \geq \epsilon^2] \;\leq\; E_\theta[(\hat{\theta}-\theta)^2]\,/\,\epsilon^2 \;=\; MSE_\theta(\hat{\theta})\,/\,\epsilon^2\,,$$

    so if $MSE_\theta(\hat{\theta}) \to 0$ then $P_\theta[|\hat{\theta}-\theta| \geq \epsilon] \to 0$. ■

- Corollary: If $\lim_{n\to\infty} \text{Bias}_\theta(\hat{\theta}) = 0$, and $\lim_{n\to\infty} \text{Var}_\theta(\hat{\theta}) = 0$, then $\hat{\theta}$ is consistent.

  - Proof: In this case,

$$\lim_{n\to\infty} MSE_\theta(\hat{\theta}) \;=\; \lim_{n\to\infty} [\text{Var}_\theta(\hat{\theta}) + (\text{Bias}_\theta(\hat{\theta}))^2] \;=\; 0\,. \quad ■$$

- Example: If $P_\theta = N(\theta, 1)$, and $\hat{\theta} = \overline{x}$, then $\text{Bias}_\theta(\hat{\theta}) = 0$, and $\text{Var}_\theta(\hat{\theta}) = 1/n \to 0$, so $\hat{\theta}$ is consistent.

  - If instead try $\hat{\theta} = x_1$, then still $\text{Bias}_\theta(\hat{\theta}) = 0$, but now $\text{Var}_\theta(\hat{\theta}) = 1 \not\to 0$. In fact, this $\hat{\theta}$ is <u>not</u> consistent since $P[|\hat{\theta}-\theta| \geq \epsilon]$ does not change with n and so does not $\to 0$.

- Referendum Example: Estimate θ by $\hat{\theta} = c/n$. Then $\text{Bias}_\theta(\hat{\theta}) = 0$, and $\text{Var}_\theta(\hat{\theta}) = \theta(1-\theta)\,/\,n \to 0$ as $n \to \infty$. So, $\hat{\theta}$ is consistent.

- For any model, if observe $x_1, \ldots, x_n$, and estimate cdf $F_\theta(z) = P_\theta(X \leq z)$ by $\hat{F}(z) = \frac{1}{n}\sum_{i=1}^n I_{(-\infty,z]}(x_i)$, then since $P_\theta(x_i \leq z) = F_\theta(z)$, it follows from the WLLN that $\hat{F}(z) \to F_\theta(z)$ in probability as $n \to \infty$, so $\hat{F}(z)$ is a consistent estimator of $F_\theta(z)$.

- Similarly, $\frac{1}{n}\sum_{i=1}^{n} I_z(x_i)$ is consistent estimator of $P_\theta(X = z)$, again by WLLN.

- CONFIDENCE INTERVALS (6.3.2):
- Example: Suppose $P_\theta = N(\theta, 1)$, and estimate θ by $\overline{x}$. How close are we?
  - Well, MSE is $1/n$. So, if e.g. $n = 16$, and $\overline{x} = 5$, then on average $(\theta - 5)^2 \approx 1/16$, so $|\theta - 5| \approx 1/4$, so perhaps θ is likely to be between 4.75 and 5.25. But how sure can we be?
  - Well, $\overline{x} \sim N(\theta,\, 1/n)$, so that $\sqrt{n}(\overline{x} - \theta) \sim N(0, 1)$, with cdf $\Phi(z)$.
  - Fact: $\Phi(-1.96) \doteq 0.025$. [Text Table D.2.] Hence, if $Z \sim N(0, 1)$, then $P(Z < -1.96) \doteq 0.025$. Similarly $P(Z > +1.96) \doteq 0.025$. So, $P(-1.96 < Z < 1.96) \doteq 0.95$. [Note: The figure 1.96 is so important that you should remember it.]
  - Thus, $P(-1.96 < \sqrt{n}(\overline{x} - \theta) < 1.96) \doteq 0.95$. So, $P(-1.96/\sqrt{n} < \overline{x} - \theta < 1.96/\sqrt{n}) \doteq 0.95$. So, $P(\overline{x} - 1.96/\sqrt{n} < \theta < \overline{x} + 1.96/\sqrt{n}) \doteq 0.95$. ["$(\overline{x} - 1.96/\sqrt{n},\ \overline{x} + 1.96/\sqrt{n})$ is 95% confidence interval for θ."]
  - e.g. $n = 16$, $\overline{x} = 5$, then $1.96/\sqrt{n} \doteq 0.49$, so $P(5 - 0.49 < \theta < 5 + 0.49) \doteq 0.95$. Roughly speaking, we're 95% sure that θ is between 4.5 and 5.5. ["19 times out of 20"]
  - Error gets smaller as $n \to \infty$. [Not surprising since $\hat{\theta}$ is consistent.]
  - If instead want to be 99% sure, then just replace "1.96" by "2.57", since $\Phi(2.57) \doteq 0.995$. [Or, if replace "1.96" by "1", then 68% sure.]
- If instead $P_\theta = N(\theta, \sigma_0^2)$ (with σ_0^2 known), then instead $\sqrt{n/\sigma_0^2}(\overline{x} - \theta) \sim N(0, 1)$, so instead $P(\overline{x} - 1.96\,\sqrt{\sigma_0^2/n} < \theta < \overline{x} + 1.96\,\sqrt{\sigma_0^2/n}) \doteq 0.95$.

—— **END WEDNESDAY 4** ——

[Test #1 from 3–5 on Wednesday Feb 11: Surnames A–Li in Medical Sciences Building (1 King's College Circle) room 3153; Surnames Ll–Z in Canadiana Gallery (14 Queen's Park Crescent, behind Sig Sam Library) room 150. No aids allowed. Bring your T-Card!]

[Lots of TA office hours [and more] available on web.]

- Exercise 6.1.18: $\Omega = \{1, 2\}$, $T(s) = f_1(s)/f_2(s)$, show T is minimal sufficient statistic. [I've gotten many questions about this ... and there are many different approaches ... but here's the most direct.] [Assume $f_i(s) > 0 \, \forall s \in S$ to avoid complications.] Note that

$$L(\theta \,|\, s_1) = K \, L(\theta \,|\, s_2) \ \forall \theta \in \Omega \ \Leftrightarrow \ L(\theta \,|\, s_1)/L(\theta \,|\, s_2) = K \ \forall \theta \in \Omega$$

$$\Leftrightarrow \ L(1 \,|\, s_1)/L(1 \,|\, s_2) = L(2 \,|\, s_1)/L(2 \,|\, s_2) \ \Leftrightarrow \ L(1 \,|\, s_1)/L(2 \,|\, s_1) = L(1 \,|\, s_2)/L(2 \,|\, s_2)$$

$$\Leftrightarrow \ T(s_1) = T(s_2) \, . \ \blacksquare$$

- [Also, don't worry too much about Exercise 6.2.14.]

Previous Class:
∗ $\hat{\theta}$ consistent if $MSE_\theta(\hat{\theta}) \to 0$.
∗ Estimation of probabilities by corresponding "fraction of data" is consistent, by WLLN (Text Thm 4.2.1).
∗ Confidence intervals.
—— Example: if $P_\theta = N(\theta, 1)$, then 95% C.I. given by $\overline{x} \pm 1.96 \, / \, \sqrt{n}$.
—— If instead $P_\theta = N(\theta, \sigma_0^2)$, then instead get $\overline{x} \pm 1.96 \, \sqrt{\sigma_0^2/n}$.

- CONFIDENCE INTERVALS, continued.

- Location-Scale Model: Suppose $\theta = (\mu, \sigma^2)$, and $P_\theta = N(\mu, \sigma^2)$, i.e. μ and σ^2 both <u>unknown</u>. Then what is 95% confidence interval for μ?

  - Well, can estimate σ^2 by S^2, so might hope that $P(\overline{x} - 1.96 \sqrt{S^2/n} < \mu < \overline{x} + 1.96 \sqrt{S^2/n}) \approx 0.95$.

  - <u>However</u>, actually the uncertainty in σ^2 requires a <u>larger</u> confidence interval.

  - Recall that $\sqrt{n}(\overline{x} - \mu)/\sigma \sim N(0, 1)$ and $(n-1)S^2/\sigma^2 \sim \chi^2(n-1)$, indep., so

$$\sqrt{n/S^2} \, (\overline{x} - \mu) \ = \ \frac{\sqrt{n}(\overline{x} - \mu)/\sigma}{\sqrt{(n-1)S^2/\sigma^2(n-1)}} \ \sim \ t(n-1) \, ,$$

a t distribution with $n-1$ degrees of freedom. Hence, if a_n is such that $P(-a_n < T_n < a_n) = 0.95$ whenever $T_n \sim t(n)$, then $P(\overline{x} - a_{n-1}\sqrt{S^2/n} < \mu < \overline{x} + a_{n-1}\sqrt{S^2/n}) \doteq 0.95$.

  - Always have $a_n > 1.96$, i.e. confidence intervals larger because of uncertainty in σ^2. However, $a_n \approx 1.96$ if n is large.
  - e.g. $a_3 = 3.18$, $a_{10} = 2.23$, $a_{50} = 2.01$. [Text Table D.4. You do not need to memorise these values.]
  - Can similarly get confidence intervals for σ^2 in terms of S^2, using $\chi^2(n-1)$ distribution.

- Example: Election poll, candidates A, B, C. Ask n people who they will vote for; c of them say A. Find confidence interval for θ = fraction of votes A will get.
  - Let $\hat{\theta} = c/n$.
  - Know $c \sim \text{Binomial}(n, \theta)$, so $E_\theta(\hat{\theta}) = \theta$, and $MSE_\theta(\hat{\theta}) = \text{Var}_\theta(\hat{\theta}) = \theta(1-\theta)\,/\,n$. But how to get confidence interval?
  - If n small, can perhaps compute with Binomial(n, θ) directly. But what if n large?
  - Use CLT! If n large, then $(\hat{\theta} - \theta)\,/\sqrt{\text{Var}_\theta(\hat{\theta})} \sim N(0,1)$, i.e. $\sqrt{n/\theta(1-\theta)}\,(\hat{\theta} - \theta) \sim N(0,1)$.
  - Hence, like above, $P(\hat{\theta} - \delta_n < \theta < \hat{\theta} + \delta_n) \approx 0.95$, where $\delta_n = 1.96\sqrt{\theta(1-\theta)/n}$ = "95% margin of error".
  - Another problem: θ unknown! Two options: (1) "Plug-In Estimate": replace θ by its estimate, $\hat{\theta}$. (2) "Conservative Option": Use that always $\theta(1-\theta) \leq 1/4$, so if $\delta_n = 1.96\sqrt{(1/4)/n} = 1.96\,/\,2\sqrt{n} = 0.98\,/\sqrt{n}$, then $P(\hat{\theta} - \delta_n < \theta < \hat{\theta} + \delta_n) \geq 0.95$. [Good, but conservative.]

- What do real polling companies do?
  - e.g. Ipsos-Reid mayor's poll, November 3, 2003 (one week before mayoral election). Phoned 700 adult Torontonians. Got estimate Miller 37%, Tory 31%, "accurate within ±3.7%, 19 times out of 20".

- Check: $0.98 \,/\, \sqrt{700} \doteq 0.03704052 \doteq 3.7\%$. i.e. polling companies usually use option (2) above.
- If instead wanted 99% certainty, then replace 1.96 by 2.57, get error $2.57 \,/\, 2\sqrt{700} \doteq 0.04856843 \doteq 4.9\%$.

- Basketball Example: Score 7 out of 10 foul shots. What is approximate 95% confidence interval for p? Here $0.98 \,/\, \sqrt{10} \doteq 0.31$, so p could be anywhere in $(0.7 - 0.31,\ 0.7 + 0.31) = (0.39,\ 1.01)$. Large interval! [Also crazy, since must have $p < 1$, i.e. $n = 10$ is too small to accurately use normal approximation.]
  - If instead score 70 out of 100, then $0.98 \,/\, \sqrt{100} = 0.098 \approx 0.1$, so 95% confidence interval for p is approx. $(0.6, 0.8)$.
  - If use Plug-In Estimate instead, then for $n = 10$ case get margin of error $= 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n} = 1.96\sqrt{0.7(0.3)/10} \doteq 0.28$, and for $n = 100$ case get margin of error $= 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n} = 1.96\sqrt{0.7(0.3)/100} \doteq 0.090$. [In both cases, margin of error a little smaller.]
  - SAMPLE SIZE calculation (6.3.4): How many shots must we observe to get 95% sure of being within, say, 0.02 of the true value of p? Want 95% margin of error ≤ 0.02, i.e. $0.98 \,/\, \sqrt{n} \leq 0.02$, i.e. $n \geq (0.98/0.02)^2 = 2401$. So, would require at least 2401 shots.

- Note: Can use this CLT in many cases. If you can find (say) C_1 and C_2 such that, under P_θ, $Z = C_1(\overline{x} - C_2)$ has mean 0 and variance 1, then for large n, $Z \sim N(0,1)$, so $P[|Z| \geq 1.96] \cdot 0.95$.
- HYPOTHESIS TESTING (6.3.3)
- "Statitus" Example: Have either fair coin or two-headed coin. Get three heads in a row. Are we sure we have two-headed coin?
- Have "null hypothesis" H_0 that coin is fair, versus "alternative hypothesis" H_1 that coin is two-headed.
- Defn: The <u>P-value</u> of an experiment, is the probability that we <u>would</u> observe that result, or a result "at least as surprising", <u>if</u> the null hypothesis H_0 is true.

- "Statitus": P-value is $(1/2)^3 = 1/8 = 0.125$. Small enough to conclude that H_0 is false??
  - No! Usually require P-value < 0.05 to conclude H_0 false. ["Three heads is not statistically significant."]
  - If instead get five heads in a row, then P-value $= (1/2)^5 = 1/32 \doteq 0.031 < 0.05$, enough to conclude that H_0 is false and we must have the two-headed coin. ["Five heads is statistically significant."]
  - Suppose we demand 99% significance instead, i.e. require P-value < 0.01. Then need seven heads in a row, to get P-value $= (1/2)^7 = 1/128 \doteq 0.008 < 0.01$.

—— **END MONDAY 5** ——

- Example: $P_\theta = N(\theta, 1)$. Suppose have hypothesis H_0: $\theta = \theta_0 = 5$ (say), compared to $H_1 : \theta \neq 5$. Then observe $x_1, \ldots, x_n$, and compute $\overline{x} = 5.1$ (say). Can we be sure that H_0 is wrong?
  - Well, here P-value is $P_5[|\overline{x} - 5| \geq 0.1]$.
  - But under P_5, $\overline{x}$ has distribution $N(5, 1/n)$, so $\sqrt{n}(\overline{x} - 5) \sim N(0, 1)$. Hence, P-value is

$$P_5[|\overline{x} - 5| \geq 0.1] = P_5[|\sqrt{n}(\overline{x} - 5)| \geq 0.1\sqrt{n}] = P[|Z| \geq 0.1\sqrt{n}]$$

$$= P[Z \leq -0.1\sqrt{n}] + P[Z \geq 0.1\sqrt{n}] = 2\,P[Z \leq -0.1\sqrt{n}] = 2\,\Phi(-0.1\sqrt{n})\,,$$

    where $Z \sim N(0, 1)$. ["Z-test"]
  - e.g. [Using text Table D.2, to be supplied if needed for tests.] $n = 1$: P-value $\doteq 0.92$; $n = 10$: P-value $\doteq 0.75$; $n = 100$: P-value $\doteq 0.32$; $n = 200$: P-value $\doteq 0.16$; $n = 400$: P-value $\doteq 0.046$; $n = 700$: P-value $\doteq 0.0082$.
  - Conclude that to distinguish between $H_0 : \theta = 5$, and $H_1 : \theta \neq 5$, when $\overline{x} = 5.1$, requires SAMPLE SIZE (Sect. 6.3.4) of about 400 at 95% level, or about 700 at 99% level.
- If instead $P_\theta = N(\theta, \sigma_0^2)$, with $\sigma_0^2 > 0$ known, then instead obtain P-value of $2\,\Phi(-|\overline{x} - \theta_0|\,\sqrt{n/\sigma_0^2})$. [Exercise!]

- Bernoulli Model (Text Example 6.3.11): Suppose again that $\Omega = [0,1]$, and $P_\theta(\text{die}) = \theta$, $P_\theta(\text{live}) = 1-\theta$. Suppose "usually" $\theta = \theta_0$ (known), but environment has changed. Question: Do we still have $\theta = \theta_0$?
  - Here $H_0 : \theta = \theta_0$, while $H_1 : \theta \neq \theta_0$.
  - Suppose observe n patients, of whom c die. Assume n large. Let $\delta = |(c/n) - \theta_0|$, observed deviation from θ_0.
  - Then P-value is $P_{\theta_0}[|(c/n) - \theta| \geq \delta]$.
  - Under P_{θ_0}, $c \sim \text{Binomial}(n, \theta_0)$, with mean $n\theta_0$ and variance $n\theta_0(1-\theta_0)$. So, $\overline{x} = c/n$ has mean θ_0 and variance $\theta_0(1-\theta_0)/n$. [Here θ_0 known, so don't need to bound variance by $1/4n$.]
  - Hence if $Z = \sqrt{n/\theta_0(1-\theta_0)}\,(\overline{x} - \theta_0)$, then for large n, $Z \sim N(0,1)$.
  - So, P-value is given by

$$P_{\theta_0}[|(c/n) - \theta| \geq \delta] = P_{\theta_0}[|Z| \geq \delta\, \sqrt{n/\theta_0(1-\theta_0)}] = 2\,\Phi\big(-\delta\, \sqrt{n/\theta_0(1-\theta_0)}\big)\,.$$

  - e.g. $\theta_0 = 0.2$, observe $n = 1000$, $c = 250$. Can we conclude the new environment is more dangerous? Here $\delta = |(250/1000) - 0.2| = 0.05$, and P-value is

$$2\,\Phi\big(-\delta\, \sqrt{n/\theta_0(1-\theta_0)}\big) = 2\,\Phi\big(-0.05\, \sqrt{1000/0.2(0.8)}\big) \doteq 2\,\Phi(-3.95) \doteq 0.000077\,.$$

    So yes, there is a (highly) statistically significant change: it's gotten more dangerous!
  - Suppose instead had $n = 4$ and $c = 1$. Then still $c/n = 0.25$, and $\delta = |(c/n) - 0.2| = 0.05$. But would the change still be statistically significant? (No!)

—— **END WEDNESDAY 5** ——

[Reminder: Test #1 on Wednesday, 3–5: Surnames A–Li in MS 3153; Surnames Ll–Z in CG 150. No aids allowed. Bring your T-Card!]

[No classes next week (Reading Week).]

Previous Class:

∗ Examples re P-values:

—— Case $P_\theta = N(\theta, 1)$, $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, P-value $= P_{\theta_0}[|\overline{X} - \theta_0| \geq \delta] = 2\,\Phi(-\delta\sqrt{n})$, where $\delta = |\overline{x} - \theta_0|$ (observed value, as opposed to random variable $\overline{X}$ in prob).

—— Case $P_\theta = N(\theta, \sigma_0^2)$, P-value $= 2\,\Phi(-\delta\sqrt{n/\sigma_0^2})$.

—— Bernoulli Model, P-value $\approx 2\,\Phi(-\delta\sqrt{n/\theta_0(1-\theta_0)})$, because of CLT (for n large).

- Bernoulli Model revisited: $\Omega = [0, 1]$, $P_\theta(\text{die}) = \theta$, $P_\theta(\text{live}) = 1 - \theta$, $H_0 : \theta = \theta_0$, observe n patients of whom c die, set $\delta = |(c/n) - \theta_0|$ (observed difference), then P-value equals

$$P_{\theta_0}[|(C/n) - \theta_0| \geq \delta] = 2\,\Phi\big(-\delta\ \sqrt{n/\theta_0(1-\theta_0)}\big)\,.$$

  - One-Sided Tests: Suppose instead that we're only worried about one "side" of the change in θ, namely θ getting larger. i.e. still $H_0 : \theta = \theta_0$, but now $H_1 : \theta > \theta_0$ instead of $H_1 : \theta \neq \theta_0$.
  - In that case, replace P-value $P_{\theta_0}[|(C/n) - \theta| \geq \delta]$ by just $P_{\theta_0}[(C/n) - \theta \geq \delta]$.
  - This change removes the factor of "2" in P-value calculation, i.e. gives P-value $= \Phi\big(-\delta\ \sqrt{n/\theta_0(1-\theta_0)}\big)$ which is half as large.
  - Whether to use Two-Sided (usual) or One-Sided test is a matter of judgement, depending on the problem. [Usually just assume Two-Sided.]

- Location-Scale Model P-values [Text Example 6.3.13]: $P_\theta = N(\mu, \sigma^2)$ with μ and σ^2 both unknown. Have hypothesis $H_0 : \mu = \mu_0$. Observe $x_1, \ldots, x_n$ with deviation $\delta = |\overline{x} - \mu_0|$. What is P-value?
  - Recall that $T \equiv \sqrt{n/S^2}\,(\overline{X} - \mu) \sim t(n-1)$. So, P-value is

$$P_\theta[|\overline{X} - \mu| \geq \delta] \ = \ P[|T| \geq \delta\,\sqrt{n/S^2}] \ = \ 2\,P[T \leq -\delta\,\sqrt{n/S^2}]\,.$$

[Can find from statistical package. Will be provided as needed on the class tests.] ["t-test"]

- e.g. $\mu_0 = 5$, $\overline{x} = 5.1$, $S^2 = 1$, $n = 100$: Get P-value equal to 0.3197, compared to 0.3173 if $\sigma^2 = 1$ is known. [i.e. P-value slightly larger due to uncertainty in σ^2.]
- If instead use one-sided test, i.e. test $H_0 : \mu = \mu_0$ versus $H_1 : \mu > \mu_0$, then remove factor of "2", get P-value equal to 0.1599.

- Statitus partial-treatment: Suppose statitus is usually 50% fatal. Company claims that with their treatment, it's "less" fatal. We observe 8 patients, of whom just 1 dies. Are we sure the company is correct?
  - Let $\Omega = [0, 1]$, $P_\theta[\text{die}] = \theta$, $P_\theta[\text{live}] = 1 - \theta$. Then $H_0 : \theta = 0.5$, and $H_1 : \theta < 0.5$. What is P-value?
  - Since $n = 8$ is small, don't use CLT. Also, since they claim it is less fatal, use one-sided test. So, P-value is $P[\le 1 \text{ die}]$.
  - Under H_0,

$$P[\le 1 \text{ die}] = P_{0.5}[\le 1 \text{ die}] = P_{0.5}[0 \text{ die}] + P_{0.5}[1 \text{ die}]$$

$$= (0.5)^8 + \binom{8}{1}(0.5)^7(0.5)^1 = 9/2^8 \doteq 0.035 < 0.05\,.$$

    So, 95% confident that treatment helps. [Not 99% confident, though!]
  - If instead just observed five patients, of whom one died, then compute [Exercise!] that P-value $= 6/2^5 \doteq 0.19$. In this case, we're not sure if it helped.

- METHOD OF MOMENTS (6.4.1)
- Another way to estimate θ is to find the value $\hat{\theta}$ such that mean of $P_{\hat{\theta}}$ equals $\overline{x}$. [And, if necessary, $E_{\hat{\theta}}[X^2] = \frac{1}{n}\sum_{i=1}^n (x_i)^2$, etc.] ["Method of Moments (MoM) Estimator"]
- Example: $P_\theta = N(\theta, 1)$. Then mean of P_θ is θ. So, for MoM Estimator, want $\hat{\theta} = \overline{x}$. [Same as MLE.]
- Example: $P_\theta = \text{Exp}(\theta)$. Then mean of P_θ is $1/\theta$. So, for MoM Estimator, want $1/\hat{\theta} = \overline{x}$, i.e. $\hat{\theta} = 1/\overline{x}$.

- Example: $P_\theta = \text{Uniform}[0, \theta]$. MLE is $\max_{1 \le i \le n} x_i$. What is MoM Estimator?
  - Well, mean of P_θ is $\theta/2$. So, must have $\hat{\theta}/2 = \overline{x}$, i.e. $\hat{\theta} = 2\overline{x}$.
  - We've seen this before! ["$\hat{\theta}_2$"] We know it's consistent, has $MSE_\theta(\hat{\theta}) \to 0$, etc.
- Example: $P_\theta = \text{Uniform}[-\theta, \theta]$. MLE is $\max_{1 \le i \le n} |x_i|$. What is MoM Estimator?
  - Here mean of P_θ is 0, which doesn't help. So, must consider <u>second</u> moment.
  - Second moment of P_θ is $(2\theta)^2/12 = \theta^2/3$. So, want $\hat{\theta}^2/3 = \frac{1}{n}\sum_{i=1}^n (x_i)^2$, i.e. $\hat{\theta} = \sqrt{(3/n)\sum_{i=1}^n (x_i)^2}$.

—— **END MONDAY 6** ——

[Held Test #1, then week off for Reading Week ...]

- SUMMARY SO FAR: Have learned basics of "classical statistics":
  - Inference when prob dist known or unknown.
  - Statistical Models, likelihood functions.
  - Maximum Likelihood Estimators, Score Equation.
  - (Minimal) Sufficient Statistics.
  - Bias, MSE, Consistency.
  - Confidence intervals, hypothesis testing.
  - Method-of-Moments estimators
- INTRODUCTION TO BAYESIAN INFERENCE (7.1)
- COIN EXAMPLE: Suppose I have either regular or two-headed coin.
  - What is probability I have two-headed coin? (Undefined?)
  - Suppose I flip it once, and get heads. <u>Now</u> what is probability I have two-headed coin? (Still undefined?)
  - In "classical" statistics, these probabilities are <u>undefined</u>. However, an alterna-

tive approach, “Bayesian statistics”, says that every unknown has probabilities associated with it.

– Bayesian statistics says start with a prior distribution of what you think at the beginning. e.g. $\Pi(\text{regular}) = \Pi(\text{two-headed}) = 1/2$.

– Then if get one head, then new probability of two-headed coin is equal to old probability, conditional on seeing one head:

$$P(\text{two-headed} \,|\, \text{head}) = \frac{P(\text{two-headed, head})}{P(\text{head})}$$

$$= \frac{P(\text{two-headed, head})}{P(\text{two-headed, head}) + P(\text{regular, head})} = \frac{(1/2)(1)}{(1/2)(1) + (1/2)(1/2)}$$

$$= \frac{1/2}{3/4} = 2/3\,.$$

– If get k heads in a row, then

$$P(\text{two-headed} \,|\, k \text{ heads}) = \frac{P(\text{two-headed, } k \text{ heads})}{P(k \text{ heads})}$$

$$= \frac{P(\text{two-headed, } k \text{ heads})}{P(\text{two-headed, } k \text{ heads}) + P(\text{regular, } k \text{ heads})}$$

$$= \frac{(1/2)(1)^k}{(1/2)(1)^k + (1/2)(1/2)^k} = \frac{1}{1 + (1/2)^k}\,.$$

This $\to 1$ as $k \to \infty$.

– Suppose instead had prior $\Pi(\text{regular}) = 1/3$, $\Pi(\text{two-headed}) = 2/3$. Then if get k heads in a row, then

$$P(\text{two-headed} \,|\, k \text{ heads}) = \frac{P(\text{two-headed, } k \text{ heads})}{P(k \text{ heads})}$$

$$= \frac{P(\text{two-headed, } k \text{ heads})}{P(\text{two-headed, } k \text{ heads}) + P(\text{regular, } k \text{ heads})}$$

$$= \frac{(2/3)(1)^k}{(2/3)(1)^k + (1/3)(1/2)^k} = \frac{2}{2 + (1/2)^k}\,.$$

This still $\to 1$ as $k \to \infty$.

- A Bayesian Model consists of Statistical Model $\{P_\theta : \theta \in \Omega\}$ together with a prior distribution Π on Ω.
  - Discrete case: Π has probability function $\pi(\theta)$ = probability that θ is true.
  - Absolutely continuous case: Π has density function $\pi(\theta)$, so that probability θ between a and b is $\int_a^b \pi(\theta)\, d\theta$.
- Then pair (θ, s) has prior joint probability (or density) function $\pi(\theta)\, f_\theta(s)$.
- Hence, prior marginal distribution for s is $m(s) = \sum_{\theta\in\Omega} \pi(\theta)\, f_\theta(s)$ [discrete case], or $m(s) = \int_{\theta\in\Omega} \pi(\theta)\, f_\theta(s)\, d\theta$ [absolutely continuous case]. "Prior Predictive Distribution"
- Then once we observe some data s, then get conditional probability (or density) function for θ:
$$\pi(\theta \mid s) = \frac{\pi(\theta)\, f_\theta(s)}{m(s)}\,.$$
"Posterior Distribution" ["Posterior equals prior times likelihood, normalised."]
- Coin Example again:
  - Here $\pi(\text{two-headed}) = \pi(\text{regular}) = 1/2$.
  - $f_{\text{two-headed}}(\text{head}) = 1$; $f_{\text{two-headed}}(\text{tail}) = 0$; $f_{\text{regular}}(\text{head}) = f_{\text{regular}}(\text{tail}) = 1/2$.
  - $m(\text{head}) = \pi(\text{two-headed})\, f_{\text{two-headed}}(\text{head}) + \pi(\text{regular})\, f_{\text{regular}}(\text{head}) = (1/2)(1) + (1/2)(1/2) = 3/4$. Also $m(\text{tail}) = \pi(\text{two-headed})\, f_{\text{two-headed}}(\text{tail})$ $+\pi(\text{regular})\, f_{\text{regular}}(\text{tail}) = (1/2)(0) + (1/2)(1/2) = 1/4$.
  - Then
$$\pi(\text{two-headed} \mid \text{head}) = \frac{\pi(\text{two-headed})\, f_{\text{two-headed}}(\text{head})}{m(\text{head})} = \frac{(1/2)\,(1)}{3/4} = 2/3\,.$$
[Same as before.]
  - Also
$$\pi(\text{regular} \mid \text{head}) = \frac{\pi(\text{regular})\, f_{\text{regular}}(\text{head})}{m(\text{head})} = \frac{(1/2)\,(1/2)}{3/4} = 1/3\,.$$
  - Also
$$\pi(\text{two-headed} \mid \text{tail}) = \frac{\pi(\text{two-headed})\, f_{\text{two-headed}}(\text{tail})}{m(\text{tail})} = \frac{(1/2)\,(0)}{3/4} = 0\,.$$

– If instead observe k heads, then $m(k \text{ heads}) = \pi(\text{two-headed})\, f_{\text{two-headed}}(k \text{ heads}) + \pi(\text{regular})\, f_{\text{regular}}(k \text{ heads}) = (1/2)(1)^k + (1/2)(1/2)^k = (1/2) + (1/2)^{k+1}$. Then

$$\pi(\text{two-headed} \,|\, k \text{ heads}) = \frac{\pi(\text{two-headed})\, f_{\text{two-headed}}(k \text{ heads})}{m(k \text{ heads})}$$

$$= \frac{(1/2)\,(1)^k}{(1/2) + (1/2)^{k+1}} = \frac{1}{1 + (1/2)^k}\,.$$

[Same as before.]

• EXAMPLE: BERNOULLI MODEL. (Text p. 354.)

– Here $S = \{0, 1\}$, $\Omega = [0, 1]$, and $P_\theta(1) = \theta$, $P_\theta(0) = 1 - \theta$. Suppose prior is UNIFORM on Ω, so that $\pi(\theta) \equiv 1$. Suppose observe $x_1, \ldots, x_n \in S$. What is posterior?

– Here $f_\theta(x_1, \ldots, x_n) = \theta^c (1 - \theta)^{n-c}$, where $c = \#\{i;\ x_i = 1\} = n\overline{x}$.

– Then $m(x_1, \ldots, x_n) = \int_{\theta \in \Omega} \pi(\theta)\, f_\theta(x_1, \ldots, x_n)\, d\theta = \int_0^1 (1) \theta^c (1-\theta)^{n-c} d\theta = \int_0^1 y^c (1-y)^{n-c} dy$. Hard! [FACT: This equals $\Gamma(c+1)\,\Gamma(n-c+1)\,/\,\Gamma(n+2)$, or $c!\,(n-c)!\,/\,(n+1)!$. But never mind that!]

– Then posterior density is given by

$$\pi(\theta \,|\, x_1, \ldots, x_n) = \frac{\pi(\theta)\, f_\theta(x_1, \ldots, x_n)}{m(x_1, \ldots, x_n)} = \frac{(1)\, \theta^c (1-\theta)^{n-c}}{\int_0^1 y^c (1-y)^{n-c} dy}\,.$$

In fact, this is a Beta distribution, $\text{Beta}(c+1, n-c+1)$. [Text pp. 60, 654.]

– Posterior provides our best understanding, given our prior Π and the data $x_1, \ldots, x_n$, of all the probabilities for θ.

– Once we have posterior, then we might estimate θ by the posterior mean estimator. Now, the mean of the $\text{Beta}(c+1, n-c+1)$ distribution is $(c+1)/[(c+1) + (n-c+1)] = (c+1)/(n+2)$. Hence, the posterior mean estimator for θ is $\hat{\theta} = (c+1)/(n+2)$. This is close to our "usual" estimator c/n, but a bit closer to $1/2$.

—— **END MONDAY 7** ——

Previous Class:

$*$ Quick review of classical statistics.

$*$ Bayesian inference:

$*$ Coin Example (two-headed or regular).

$*$ Prior distribution Π, with prob/dens fn $\pi(\theta)$.

$*$ Prior predictive distribution $m(s) = \sum_{\theta \in \Omega} \pi(\theta) f_\theta(s)$.

$*$ Posterior prob/dens fn $\pi(\theta \,|\, s) = \pi(\theta) f_\theta(s) \,/\, m(s)$.

$*$ Bernoulli Model: $\pi(\theta) = 1$ (Uniform), then $\pi(\theta \,|\, x_1, \ldots, x_n) \propto \theta^c (1-\theta)^{n-c}$, i.e. $\Pi(\theta \,|\, x_1, \ldots, x_n) =$ Beta$(c+1, n-c+1)$.

—— Then can e.g. estimate θ by posterior mean $\hat{\theta} = (c+1)/(n+2)$.

- Note that the variance of the Beta$(c+1, n-c+1)$ distribution is $(c+1)(n-c+1) \,/\, (n+3)(n+2)^2$, and this provides a measure of how uncertain we are about the estimate $(c+1)/(n+2)$. As $n \to \infty$, since $0 \le c \le n$, we see that variance $\to 0$, i.e. we're more and more sure.

• Consider again Bernoulli model, but this time with prior density $\pi(\theta) = 4\,\theta^3$ for $\theta \in \Omega \equiv [0,1]$. [i.e. we think it's more likely that θ is larger]

- Still have $f_\theta(x_1, \ldots, x_n) = \theta^c (1-\theta)^{n-c}$, where $c = \#\{i;\ x_i = 1\} = n\overline{x}$.

- $m(x_1, \ldots, x_n)$ still hard to compute.

- Posterior density is given by

$$\pi(\theta \,|\, x_1, \ldots, x_n) = \frac{\pi(\theta)\, f_\theta(x_1, \ldots, x_n)}{m(x_1, \ldots, x_n)} = \frac{4(\theta^3)\, \theta^c (1-\theta)^{n-c}}{m(x_1, \ldots, x_n)} = \frac{4\theta^{c+3} (1-\theta)^{n-c}}{m(x_1, \ldots, x_n)}\,.$$

- We observe that this is a Beta$(c+4, n-c+1)$ distribution. [Text pp. 60, 654.] (Don't need to bother computing normalisation constants.)

- Posterior mean equals $(c+4) \,/\, [(c+4)+(n-c+1)] = (c+4) \,/\, (n+5)$. ["$a/(a+b)$"] A bit larger than previous posterior mean of $(c+1) \,/\, (n+2)$.

• LOCATION NORMAL MODEL. Suppose $S = \Omega = \mathbf{R}$, and $P_\theta = N(\theta, 1)$. Suppose prior is $\Pi = N(\mu_0, \tau_0^2)$ for some fixed, known μ_0 and τ_0^2. Suppose we observe $x_1, \ldots, x_n$. Then we know (from before) that

$$f_\theta(x_1, \ldots, x_n) = K \exp\big(-\frac{n}{2}(\overline{x} - \theta)^2\big)$$

Also here

$$\pi(\theta) = \frac{1}{\sqrt{2\pi\tau_0^2}} \exp\big(-(\theta-\mu_0)^2/2\tau_0^2\big)\,.$$

Also $m(x_1,\ldots,x_n) = \int \pi(\theta)\, f_\theta(x_1,\ldots,x_n)\, d\theta$. [Don't worry about this for now.]

- Then

$$\pi(\theta\,|\,x_1,\ldots,x_n) = \frac{\pi(\theta)\, f_\theta(x_1,\ldots,x_n)}{m(x_1,\ldots,x_n)}$$

$$= \frac{\exp\big(-(\theta-\mu_0)^2/2\tau_0^2\big)\ K\ \exp\big(-\frac{n}{2}(\overline{x}-\theta)^2\big)}{\sqrt{2\pi\tau_0^2}\ m(x_1,\ldots,x_n)}\,.$$

- We compute (text pp. 355–356) that this is the density of a normal distribution with mean $((\mu_0/\tau_0^2)+n\overline{x})/((1/\tau_0^2)+n)$, and variance $1\,/\,((1/\tau_0^2)+n)$.

- Hence, posterior mean estimator is $\hat{\theta} = ((\mu_0/\tau_0^2)+n\overline{x})/((1/\tau_0^2)+n)$.

- Note that $\hat{\theta}$ is a weighted average of prior mean μ_0, and sample mean $\overline{x}$. As $n\to\infty$, $\hat{\theta}\to\overline{x}$. ("The data swamps the prior.")

- SUMMARY OF BAYESIAN STATISTICS:
  - Adds new information, the "prior distribution", to the model.
  - Then can compute a "posterior distribution" which gives a full probability distribution (not just estimate) for the unknown θ.
  - Can then e.g. estimate θ by the posterior mean.
  - Advantages: Get full distribution for θ, so can estimate probabilities, etc. Also, can encorporate "prior information", e.g. if experts "believe" certain things.
  - Disadvantages: Computations can get difficult, even for simple models. [Though not too difficult for simple <u>discrete</u> models, like Coin Example. For harder examples, entire subject of "Markov Chain Monte Carlo algorithms" devoted to trying to do computations!] Also, result depends on prior and so is perhaps "subjective".
  - Very controversial: Some statisticians are die-hard Bayesians, others are anti-Bayesian!

—— **END WEDNESDAY 7** ——

Previous Class:

$*$ More examples re Bayesian inference and posterior distributions.

- MODEL CHECKING (9.1).
- A statistical model $\{P_\theta : \theta \in \Omega\}$ is just a model; how do we know if it's appropriate?
- We hope the data approximately fits some P_θ, but we don't know which one; how to check?
- Idea: Find some statistic (i.e. function of the data) which is ancilliary, i.e. whose distribution does not depend on θ. Then see if that statistic approximately follows its distribution.
- Example: $\Omega = S = \mathbf{R}$, and $P_\theta = N(\theta, 1)$.
  - Then $X_i \sim N(\theta, 1)$, which depends on θ – not ancilliary.
  - Also $\overline{X} \sim N(\theta,\, 1/n)$, which depends on θ – not ancilliary.
  - But $(n-1)S^2 \sim \chi^2(n-1)$ which does not depend on θ – ancilliary! So can check $(n-1)s^2 = \sum_{i=1}^n (x_i - \overline{x})^2$ to see if its value is "reasonable" for the $\chi^2(n-1)$ distribution.
  - e.g. suppose $n = 101$, then $\sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(100)$. Hence $E[\sum_{i=1}^n (X_i - \overline{X})^2] = 100$, and in fact $P[74.22 < \sum_{i=1}^n (X_i - \overline{X})^2 < 129.56] \doteq 0.95$. So, if $\sum_{i=1}^n (x_i - \overline{x})^2 < 74.22$, or $\sum_{i=1}^n (x_i - \overline{x})^2 > 129.56]$, then perhaps have incorrect model.
- If instead $P_\theta = N(\theta,\, \sigma_0^2)$ with σ_0^2 known, then instead $\frac{n-1}{\sigma_0^2}\, S^2 = \frac{1}{\sigma_0^2} \sum_{i=1}^n (X_i - \overline{X})^2 \sim \chi^2(n-1)$, so use this value instead.
- However, if $P_{(\mu,\sigma^2)} = N(\mu,\, \sigma^2)$ [both unknown], then more complicated! Requires simulation to approximate. [See text Example 9.1.2.]
- Example: Suppose $\Omega = S = \mathbf{R}$, with $P_\theta = \text{Uniform}[\theta - 3,\ \theta + 3]$.
  - Then under P_θ, $X_i - \theta \sim \text{Uniform}[-3, 3]$, which does not depend on θ, however it is not a statistic [depends on unknown value, θ].
  - On the other hand, $(X_i - \theta) - (X_j - \theta) = X_i - X_j$ is an ancilliary statistic.

- Hence, so is $D = \max_{i,j}(X_i - X_j) = (\max_i X_i) - (\min_j X_j)$. ["discrepancy statistic"]
- Precise distribution of D is tricky. However, if $D > 6$ then model must be wrong. Also, for large n, expect $D \approx 6$, otherwise model wrong.

- Example: $S = \Omega = (0, \infty)$, with $P_\theta(\{\theta\}) = 4/5$ and $P_\theta(\{2\theta\}) = 1/5$. Observe $x_1, \ldots, x_n$.
  - Let $D_i = X_{i+1}/X_i$ $(1 \le i \le n-1)$.
  - Then $P_\theta[D_i = 1] = P_\theta[X_i = X_{i+1}] = (4/5)^2 + (1/5)^2 = 17/25$. Also $P_\theta[D_i = 2] = P_\theta[D_i = 1/2] = (4/5)(1/5) = 4/25$.
  - Hence, D_i is an ancilliary statistic.

- Example: $S = \Omega = (0, \infty)$, with $P_\theta = \text{Exp}(\theta)$. Observe $x_1, \ldots, x_n$. What is a good ancilliary statistic?
  - Claim: $D_i = X_{i+1} \,/\, X_i$ is ancilliary $(1 \le i \le n-1)$.
  - Proof #1: Use multivariable change-of-variable formula (text Theorem 2.9.2) to get exact distribution of D_i, and observe that it does not depend on θ. ∎
  - Proof #2: Can write $X_i = Y_i \,/\, \theta$, where $Y_i \sim \text{Exp}(1)$. Then $D_i = (Y_i/\theta)/(Y_{i+1}/\theta) = Y_i/Y_{i+1}$ whose distributions do not depend on θ. ∎

- CHI-SQUARED GOODNESS OF FIT TEST (9.1.2)

- Suppose election has k candidates, $\{1, 2, \ldots, k\}$. Suppose we think that candidate i has support p_i, so $p_1 + \ldots + p_k = 1$. We then observe preferences $x_1, \ldots, x_n$, and let $c_i = \#\{j : x_j = i\}$ be count data. (So $c_1 + \ldots + c_k = n$.)
  - If we're right about the values of p_i, then should have $(C_1, \ldots, C_k) \sim \text{Multinomial}(n, p_1, \ldots, p_k)$. How to test this?
  - Well, C_i would have mean np_i, and variance $np_i(1 - p_i)$. So, for large n, should have

$$R_i \;\equiv\; \frac{C_i - np_i}{\sqrt{np_i(1-p_i)}} \approx N(0,1)\,.$$

["i'th residual"] Ancillary statistic (approx.).

– How to combine them? Intuition: for large n, $\sum_i (R_i)^2 = \sum_i (C_i - np_i)^2/np_i(1-p_i) \sim \chi^2(k)$. Not quite due to restriction $C_1 + \ldots + C_k = n$. Instead, $X^2 \equiv \sum_i (C_i - np_i)^2/np_i \approx \chi^2(k-1)$. ["Chi-squared statistic"]

– Observed value is $x^2 \equiv \sum_i (c_i - np_i)^2/np_i$. [Text: X_0^2.]

– Then P-value is $P[X^2 \geq x^2]$, where $X^2 \sim \chi^2(k-1)$. [One-sided test, since only concerned if too far <u>off</u>.]

- Example: Three candidates 1, 2, 3. We think $p_1 = 0.6$, $p_2 = 0.3$, $p_3 = 0.1$. We then poll $n = 100$ people, and observe counts $c_1 = 45$, $c_2 = 40$, $c_3 = 15$. What is P-value?

  – Here
$$x^2 = \frac{(45-60)^2}{60} + \frac{(40-30)^2}{30} + \frac{(15-10)^2}{10} = 115/12 \doteq 9.58\,.$$

  – Also if $X^2 \sim \chi^2(2)$, then $P[X^2 \geq 9.58] \doteq 0.0083$. Small! So, we conclude that our p_i values are <u>wrong</u>.

—— **END MONDAY 8** ——

Previous Class:

∗ Model Testing:

—— Can use <u>ancilliary statistic</u> to see if model is appropriate.

—— e.g. $P_\theta = N(\theta, \sigma_0^2)$, use $\frac{n-1}{\sigma_0^2} S^2 \sim \chi^2(n-1)$.

—— Other examples: Uniform, Discrete, Exponential.

∗ Chi-Squared Goodness of Fit Test

—— THM: If $(C_1, \ldots, C_k) \sim \text{Multinomial}(n, p_1, \ldots, p_k)$, then $X^2 \equiv \sum_{i=1}^k (C_i - np_i)^2/np_i \approx \chi^2(k-1)$.

—— Proof: See e.g. Theory of Statistics, by M.J. Schervish, pages 461–462. Uses matrix analysis and normal distribution theory.

—— This gives P-value $P[X^2 \geq x^2]$ for hypothesis that $\{p_i\}$ are correct.

- Chi-Squared Goodness of Fit Test can also be used for CONTINUOUS data, by first breaking it up into discrete regions.

- Example: Suppose we think the true distribution is Exp(1), and we observe values

$x_1, \ldots, x_{100}$.

- Suppose we break up $[0, \infty)$ into, say, the intervals $I_1 = [0, 1]$, $I_2 = (1, 2]$, $I_3 = (2, 5]$, and $I_4 = (5, \infty)$. Let $C_i = \#\{j : X_j \in I_i\}$ for $i = 1, 2, 3, 4$.
- Then $P(X_j \in I_1) = \int_0^1 e^{-x} dx = 1 - e^{-1} \doteq 0.632$. $P(X_j \in I_2) = \int_1^2 e^{-x} dx = e^{-1} - e^{-2} \doteq 0.232$. $P(X_j \in I_3) = \int_2^5 e^{-x} dx = e^{-2} - e^{-5} \doteq 0.129$. $P(X_j \in I_4) = \int_5^\infty e^{-x} dx = e^{-5} \doteq 0.007$.
- Then should have $(C_1, C_2, C_3, C_4) \sim \text{Multinomial}(100, 0.632, 0.232, 0.129, 0.007)$.
- Suppose we observe $c_1 = 60$, $c_2 = 25$, $c_3 = 14$, $c_4 = 1$. Then

$$x^2 = \frac{(60 - 63.2)^2}{63.2} + \frac{(25 - 23.2)^2}{23.2} + \frac{(14 - 12.9)^2}{12.9} + \frac{(1 - 0.7)^2}{0.7} \doteq 0.524\,.$$

- If $X^2 \sim \chi^2(3)$, then $P[X^2 \geq 0.524] \doteq 0.914$. Big! So, no evidence against assumption that $X_i \sim \text{Exp}(1)$.
- Comment: Here $np_4 = 0.7$ is quite small, so test is very sensitive to value of c_4. Best to have np_i "not too small" (say, ≥ 1, or ≥ 5) if possible.

• RELATIONSHIPS AMONG VARIABLES (Chapter 10)

- Given various quantities X_i and Y_i, are they related, i.e. does the distribution of one depend on the value of the other, or not? [Equivalently: Are they dependent or independent?]

• CATAGORICAL RESPONSE MODELS (Section 10.2.1).

• Suppose we take a survey of 100 U of T graduates, and find the following count data $\{c_{ij}\}$:

	Doctor	Lawyer	Scientist	Unemployed
Have taken STA261	23	13	15	5
Have NOT taken STA261	12	10	8	14

• Question: Does taking STA261 have effect on your future?

• Here have predictor variable $X \in \{\text{Taken, Not}\}$. Also outcome variable $Y \in \{\text{Doctor, Lawyer, Scientist, Unemployed}\}$. Are they dependent or independent?

- Let $\theta_{ij} = P[X = i,\, Y = j]$; $\theta_{i\cdot} = P[X = i] = \sum_j \theta_{ij}$; $\theta_{\cdot j} = P[Y = j] = \sum_i \theta_{ij}$.
  - Then X, Y independent iff $\theta_{ij} = \theta_{i\cdot}\theta_{\cdot j}$ for all i, j. Is it true? How to test?
- If we <u>knew</u> values of $\theta_{i\cdot} = q_i$ and $\theta_{\cdot j} = r_j$, then could use chi-squared statistic

$$X^2 = \sum_{i,j} \frac{(C_{ij} - nq_i r_j)^2}{nq_i r_j} \sim \chi^2(2 \cdot 4 - 1) = \chi^2(7)\,,$$

  and do usual chi-squared test.
- But here $\theta_{i\cdot}$ and $\theta_{\cdot j}$ are unknown!
- Instead, could substitute MLE: $q_i = \frac{1}{n}\sum_j c_{ij} \equiv \frac{1}{n}c_{i\cdot}$, $r_j = \frac{1}{n}\sum_i c_{ij} \equiv \frac{1}{n}c_{\cdot j}$. But then q_i and r_j depend on the data $\{c_{ij}\}$. How does this affect the distribution?
- THEOREM (e.g. Schervish, pages 463–467): For large n,

$$X^2 = \sum_{i,j} \frac{(C_{ij} - C_{i\cdot}C_{\cdot j}/n)^2}{C_{i\cdot}C_{\cdot j}/n} \sim \chi^2((2-1)(4-1)) = \chi^2(3)\,.$$

  In general, if a catagories for X, and b catagories for Y, then $X^2 \sim \chi^2((a-1)(b-1))$.
  - This is because $(a-1)(b-1) = [ab-1] - [(a-1)+(b-1)] =$ "$k-1$"−"dim(Ω)".
- Using this, can compute P-value for no relationship, as $P[X^2 \geq x^2]$, where $X^2 \sim \chi^2((a-1)(b-1))$, and x^2 is the observed value of X^2.

——— **END WEDNESDAY 9** ———

Previous Class:

∗ Applying chi-squared test to continuous data, by "partitioning".

∗ Suppose have predictor variable $X \in \{1, \ldots, a\}$ [e.g. {Taken STA261, Not Taken}], and response variable $Y \in \{1, \ldots, b\}$ [e.g. {Doctor, Lawyer, Scientist, Unemployed}.

——— Are the variables X and Y "related", i.e. dependent?

——— Null hypothesis: independent, i.e. $P[X = i,\, Y = j] \equiv \theta_{ij} = \theta_{i\cdot}\theta{\cdot}j$.

——— Use χ^2 statistic X^2, replacing np_i by $n(C_{i\cdot}/n)(C_{\cdot j}/n) = C_{i\cdot}C_{\cdot j}/n$, i.e.

$$X^2 = \sum_{i,j} \frac{(C_{ij} - C_{i\cdot}C_{\cdot j}/n)^2}{C_{i\cdot}C_{\cdot j}/n}\,.$$

∗ THEOREM: For large n, $X^2 \approx \chi^2((a-1)(b-1))$.

—— This is because $(a-1)(b-1) = [ab-1] - [(a-1)+(b-1)] =$ "$k-1$"–"$\dim(\Omega)$"; see Text Theorem 9.1.2.

- Back to "U of T graduates" example:

	Doctor	Lawyer	Scientist	Unemployed
Have taken STA261	23	13	15	5
Have NOT taken STA261	12	10	8	14

- In this example, the observed value is

$$x^2 = \frac{(23-(35)(56)/100)^2}{(35)(56)/100} + \frac{(13-(23)(56)/100)^2}{(23)(56)/100} + \frac{(15-(23)(56)/100)^2}{(23)(56)/100} + \frac{(5-(19)(56)/100)^2}{(19)(56)/100}$$
$$+ \frac{(12-(35)(44)/100)^2}{(35)(44)/100} + \frac{(10-(23)(44)/100)^2}{(23)(44)/100} + \frac{(8-(23)(44)/100)^2}{(23)(44)/100} + \frac{(14-(19)(44)/100)^2}{(19)(44)/100}$$
$$\doteq 8.93$$

  But we expect $X^2 \sim \chi^2((4-1)(2-1)) = \chi^2(3)$. Now, if $X^2 \sim \chi^2(3)$, then $P[X^2 \geq 8.93] \doteq 0.030$. So, P-value is 0.030 – small!

- Conclusion: Taking STA261 has a significant effect on your future!

- LEAST SQUARES ESTIMATES (10.3.1):

- Unconditioned case: Suppose want to estimate $E(Y)$ based on a sample $y_1, y_2, \ldots, y_n$.

- Least Squares Principle: Estimate $E(Y)$ by $\hat{e}$, chosen to minimise $SE \equiv \sum_{i=1}^n (y_i - e)^2$.

  – Well, $\frac{\partial}{\partial e} SE = -\sum_{i=1}^n 2(y_i - e)$ [differentiable everywhere], which equals 0 iff $e = \overline{y}$.

  – Also, $(\frac{\partial}{\partial e})^2 SE = \sum_{i=1}^n 2 = 2n > 0$.

  – So, <u>if</u> all values in $\mathbf{R}$ are possible for e, then must have $\hat{e} = \overline{y}$. [Makes sense.]

  – On the other hand, if only <u>certain</u> values possible for e, then $\hat{e}$ is the possible value of e which is <u>closest</u> to $\overline{y}$. [See Text Example 10.3.1.]

- What if Y <u>depends</u> on some other variable X?

  – Need to assume some <u>model</u> for the dependence.

- LINEAR REGRESSION (10.3.2):

- Suppose X, Y related variables, and we assume $E(Y \mid X = x) = \beta_1 + \beta_2 x$ for some unknown β_1, β_2, and we observe independent draws $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. How to estimate β_1 and β_2?
  - Example: x_i = grade in STA261, y_i = salary when you graduate. Are they related? How? Is β_2 zero, or positive, or negative??
- Principle of Least Squares says choose β_1, β_2 to minimise $SE \equiv \sum_{i=1}^{n} (y_i - \beta_1 - \beta_2 x_i)^2$. How? [DRAW GRAPH.]
  - Well, SE differentiable everywhere, and $\to \infty$ as $\beta_1, \beta_2 \to \pm\infty$. Hence, minimising value must be critical point (if unique).
  - Hence, want to solve $\frac{\partial}{\partial \beta_1} SE = \frac{\partial}{\partial \beta_2} SE = 0$.
  - $\frac{\partial}{\partial \beta_1} SE = -\sum_i 2(y_i - \beta_1 - \beta_2 x_i)$, which equals 0 iff $\beta_1 = \overline{y} - \beta_2 \overline{x}$.
  - $\frac{\partial}{\partial \beta_2} SE = -\sum_i 2x_i(y_i - \beta_1 - \beta_2 x_i)$. Substituting in $\beta_1 = \overline{y} - \beta_2 \overline{x}$, we see this equals $-\sum_i 2x_i(y_i - \overline{y} - \beta_2(x_i - \overline{x}))$.
  - This equals 0 iff $\beta_2 = \sum_i x_i(y_i - \overline{y}) \,/\, \sum_i x_i(x_i - \overline{x})$.
  - Since $\sum_i \overline{x}(y_i - \overline{y}) = 0 = \sum_i \overline{x}(x_i - \overline{x})$, this is the same as $\beta_2 = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \,/\, \sum_i (x_i - \overline{x})^2 \equiv b_2$.
  - Then $\beta_1 = \overline{y} - b_2 \overline{x} \equiv b_1$.
  - Thus, (b_1, b_2) is the least-squares estimate of (β_1, β_2).
- Then the line $y = b_1 + b_2 x$ is the "line of best fit" of the data $\{(x_i, y_i)\}$. Also, $b_1 + b_2 x$ is the least-squares estimate of $E(Y \mid X = x)$. [Draw graph.] [See e.g. Text Figure 10.3.4.]
  - Can also use $b_1 + b_2 x$ to estimate the actual value of Y, given $X = x$.
  - If the x_i are all equal, then $\sum_i (x_i - \overline{x})^2 = 0$, so b_2 is <u>undefined</u>. [Makes sense since then cannot determine how $E(Y \mid X = x)$ varies with x.]

—— **END MONDAY 9** ——

Previous Class:

$*$ Chi-Squared test for Catagorical Response Models:

—— $X^2 = \sum_{i,j} \frac{(C_{ij} - C_{i\cdot}C_{\cdot j}/n)^2}{C_{i\cdot}C_{\cdot j}/n} \approx \chi^2((a-1)(b-1))$.

—— Then P-value against independence is $P[X^2 \geq x^2]$.

—— Example with $a = 2$, $b = 4$.

$*$ Least Squares Principle.

—— Unconditioned case: Estimate $E[Y]$ by $\overline{y}$, or the possible value which is closest to $\overline{y}$.

$*$ Linear Regression:

—— If $E[Y \mid X = x] = \beta_1 + \beta_2 x$, then estimate β_2 by $b_2 = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \,/\, \sum_i (x_i - \overline{x})^2$, and β_1 by $b_1 = \overline{y} - b_2 \overline{x}$.

- Are these estimators <u>unbiased</u>? That is, suppose $E(Y|X = x) = \beta_1 + \beta_2 x$, with β_1 and β_2 unknown. We observe $(x_1, y_1), \ldots, (x_n, y_n)$, and estimate (β_1, β_2) by (b_1, b_2) as above. Does $E(B_i) = \beta_i$?

  - Hard to compute $E(B_2)$, since involves $E(XY)$, etc.

  - Trick: Compute <u>conditional</u> probability, $E(B_i \mid X_1 = x_1, \ldots, X_n = x_n)$:

$$E(B_2 \mid X_1 = x_1, \ldots, X_n = x_n)$$

$$= E\left(\frac{\sum_i (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_i (X_i - \overline{X})^2} \mid X_1 = x_1, \ldots, X_n = x_n\right)$$

$$= \frac{\sum_i (x_i - \overline{x})[(\beta_1 + \beta_2 x_i) - (\beta_1 + \beta_2 \overline{x})]}{\sum_i (x_i - \overline{x})^2}$$

$$= \frac{\sum_i (x_i - \overline{x}) \beta_2 (x_i - \overline{x})}{\sum_i (x_i - \overline{x})^2} = \beta_2\,.$$

  - Then by double-expectation formula (Text Theorem 3.5.2), $E(B_2) = E[E(B_2 \mid X_1, \ldots, X_n)] = E[\beta_2] = \beta_2$. Unbiased!

  - Then $E(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = E(\overline{Y} - B_2 \overline{X} \mid X_1 = x_1, \ldots, X_n = x_n) = (\beta_1 + \beta_2 \overline{x}) - \beta_2 \overline{x} = \beta_1$. Hence, $E(B_1) = E[E(B_1 \mid X_1, \ldots, X_n)] = E[\beta_1] = \beta_1$. Also unbiased!

- What about UNCERTAINTY in estimates b_1, b_2?

– Text Theorem 10.3.3: If $E(Y \mid X = x) = \beta_1 + \beta_2 x$, and $\mathrm{Var}(Y \mid X = x) = \sigma^2$ for all $x \in \mathbf{R}$, then

$$\mathrm{Var}(B_2 \mid X_1 = x_1, \ldots, X_n = x_n) = \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2}\,,$$

$$\mathrm{Var}(B_1 \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(\frac{1}{n} + \frac{(\overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)\,,$$

$$\mathrm{Var}(B_1 + B_2 x \mid X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right)\,.$$

As $n \to \infty$, $\sum_i (x_i - \overline{x})^2 \approx n\,\mathrm{Var}(X) \to \infty$, provided $Var(X) > 0$, so all these variances $\to 0$. Hence, in this case [technically, using Text Theorem 3.5.6 to remove the conditioning], the biases are zero, and the variances $\to 0$, so the $MSE \to 0$, so the estimates are <u>consistent</u> (as well as being unbiased).

—— **END WEDNESDAY 9** ——

[Test #2 from 3–5 on Wednesday March 24, in Canadiana Gallery (14 Queen's Park Crescent, behind Sig Sam Library). Surnames A–Li in room 150, surnames Ll–Z in room 250. No aids allowed. Bring your T-Card!]

[Test #2 will cover everything covered in lectures up to the end of this week, with emphasis on material <u>not</u> covered on Test #1.]

[More TA office hours posted on web site.]

Previous Class:

∗ Linear Regression Model: $E[Y \mid X = x] = \beta_1 + \beta_2 x$.

—— Here β_1, β_2 are <u>true</u> (unknown) values. [Analogous to σ^2, etc.]

∗ Then least squares estimate for β_2 is $b_2 = \sum_i (x_i - \overline{x})(y_i - \overline{y}) \,/\, \sum_i (x_i - \overline{x})^2$; and for β_1 is $b_1 = \overline{y} - b_2 \overline{x}$.

—— Here b_1, b_2 are observed values of estimators, depending on the observed values $(x_1, y_1), \ldots, (x_n, y_n)$. [Analogous to s^2, etc.]

∗ Considered sampling properties of $B_2 = \sum_i (X_i - \overline{X})(Y_i - \overline{Y}) \,/\, \sum_i (X_i - \overline{X})^2$, and $B_1 = \overline{Y} - b_2 \overline{X}$.

—— Here B_1, B_2 are the estimators, viewed as random variables depending on the random data $(X_1, Y_1), \ldots, (X_n, Y_n)$. [Analogous to S^2, etc.]

∗ Proved that B_1, B_2 are <u>unbiased</u>, i.e. $E[B_1] = \beta_1$ and $E[B_2] = \beta_2$. Good.

—— Used trick: First computed conditional expectation, conditional on $X_1 = x_1, \ldots, X_n = x_n$. Then used double-expectation formula.
$*$ Also showed (using theorem from text) that variances $\to 0$, so that $MSE \to 0$, and estimators are consistent. Good.
—— Aside: Formally, we described conditional variances $\mathrm{Var}[B_i \,|\, X_1 = x_1, \ldots, X_n = x_n]$. Then can recover usual (unconditional) variances, $\mathrm{Var}[B_i]$, using Text Theorem 3.5.6.

- If $E(Y \,|\, X = x) = \beta_1 + \beta_2 x$ for all $x \in \mathbf{R}$, then $E[B_1 + B_2 x \,|\, X_1 = x_1, \ldots, X_n = x_n] = \beta_1 + \beta_2 x$. Hence $E[B_1 + B_2 x] = \beta_1 + \beta_2 x = E[Y \,|\, X = x]$. Thus, $B_1 + B_2 x$ is an unbiased estimator of $E[Y \,|\, X = x]$. [For interpolation / extrapolation.]

- Can also compute (Text Corollary 10.3.1) that if $E(Y \,|\, X = x) = \beta_1 + \beta_2 x$, and $\mathrm{Var}(Y \,|\, X = x) = \sigma^2$ for all $x \in \mathbf{R}$, then

$$\mathrm{Var}(B_1 + B_2 x \,|\, X_1 = x_1, \ldots, X_n = x_n) = \sigma^2 \left(\frac{1}{n} + \frac{(x - \overline{x})^2}{\sum_i (x_i - \overline{x})^2} \right).$$

  This is the MSE when estimating $E[Y \,|\, X = x]$ by $B_1 + B_2 x$ (since unbiased).

  - The square-root of this MSE is then the "standard error" of estimating $E[Y \,|\, X = x]$ by $B_1 + B_2 x$. [Don't need to memorise formula, but need it for homework.]

  - Assuming $\mathrm{Var}(X) > 0$, this $MSE \to 0$ as $n \to \infty$. Thus, $B_1 + B_2 x$ is a consistent (and unbiased) estimator of $E[Y \,|\, X = x]$.

- What if σ^2 is unknown? Can estimate σ^2 by

$$s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - b_1 - b_2 x_i)^2 .$$

  [Don't need to subtract any mean, since already $E[Y - B_1 - B_2 X] = 0$.]

  - Text Theorem 10.3.4: If $E(Y \,|\, X = x) = \beta_1 + \beta_2 x$, and $\mathrm{Var}(Y \,|\, X = x) = \sigma^2$ for all $x \in \mathbf{R}$, then $E[S^2 \,|\, X_1 = x_1, \ldots, X_n = x_n] = \sigma^2$, and $E(S^2) = \sigma^2$. [Unbiased estimator.]

  - Analagous to $\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2$.

  - Intuition: We got to choose b_1, b_2, so that reduces "dimension" by 2, from n to $n - 2$. [Under additional assumptions (later), $(n-2)S^2/\sigma^2 \sim \chi^2(n-2)$.]

- How to test if X and Y are related? [e.g. does grade in STA261 really affect future income? does age really affect blood pressure?]

  - They're unrelated (actually "uncorrelated") iff $\beta_2 = 0$.

  - Our estimate b_2 may be "close" to 0. How close does it have to be? Is $b_2 = 0.1$ small enough? How to test? P-value?

  - Trick: Let
$$F = \frac{(B_2)^2 \sum_i (X_i - \overline{X})^2}{S^2} \, .$$
    ["F statistic"]

  - Why? Well, we know $E[S^2 \,|\, X_1 = x_1, \ldots, X_n = x_n] = \sigma^2$. Also
$$E[(B_2)^2 \,|\, X_1 = x_1, \ldots, X_n = x_n]$$
$$= E[B_2 \,|\, X_1 = x_1, \ldots, X_n = x_n]^2 + \mathrm{Var}[B_2 \,|\, X_1 = x_1, \ldots, X_n = x_n]$$
$$= (\beta_2)^2 + \frac{\sigma^2}{\sum_i (x_i - \overline{x})^2} \, .$$
    Hence, $E[(B_2)^2 \sum_i (X_i - \overline{X})^2 \,|\, X_1 = x_1, \ldots, X_n = x_n] = (\beta_2)^2 \sum_i (x_i - \overline{x})^2 + \sigma^2$.

  - Conclusion: <u>If</u> $\beta_2 = 0$, then $E[(B_2)^2 \sum_i (x_i - \overline{x})^2 \,|\, X_1 = x_1, \ldots, X_n = x_n] = \sigma^2$, in which case $F \approx 1$. But if F <u>large</u>, then probably $\beta_2 \neq 0$. [How large?? P-value?? More later.]

- ANOVA ("Analysis of Variance"):

- THEOREM (Text Lemma 10.3.1): If observe $(x_1, y_1), \ldots, (x_n, y_n)$, and if b_1, b_2 are linear regression coefficients as above, then
$$\sum_{i=1}^n (y_i - \overline{y})^2 = (b_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2 + \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2 \equiv RSS + ESS \, ,$$
  where RSS = regression sum of squares = amount of variation of the $\{y_i\}$ due to variation in the $\{x_i\}$, and ESS = error sum of squares = amount of variation of the $\{y_i\}$ due to deviations from the model $Y = b_1 + b_2 X$ (due to randomness in Y so that $Y \neq E[Y \,|\, X]$, and/or deviations from the model so that $E[Y \,|\, X] \neq b_1 + b_2 X$).

  - Thus, our F statistic equals $RSS/[ESS/(n-2)]$. [Distribution??]

– Also, $S^2 = ESS/(n-2)$.

—— **END MONDAY 10** ——

[Reminder: Test #2 is 3–5 next Wednesday. Surnames A–Li in room CG 150, surnames Ll–Z in room CG 250. No aids allowed. Bring your T-Card!]

Previous Class:

∗ Linear Regression Model, with $E[Y \mid X = x] = \beta_1 + \beta_2 x$, and $\mathrm{Var}[Y \mid X = x] = \sigma^2$.

—— $B_1 + B_2 x$ is unbiased, consistent estimate of $\beta_1 + \beta_2 x \equiv E[Y \mid X = x]$.

—— $S^2 \equiv \frac{1}{n-2} \sum_{i=1}^n (Y_i - B_1 - B_2 X_i)^2$ is unbiased estimator of σ^2.

—— If $F = (B_2)^2 \sum_{i=1}^n (X_i - \overline{X})^2 \,/\, S^2$, then $F \approx 1$ if $\beta_2 \approx 0$, while $F \gg 1$ if β_2 far from 0.

∗ ANOVA: $\sum_{i=1}^n (y_i - \overline{y})^2 = RSS + ESS$, where $RSS \equiv (b_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2$ and $ESS \equiv \sum_{i=1}^n (y_i - b_1 - b_2 x_i)^2$.

—— Thus, $S^2 = ESS/(n-2)$, and $F = RSS/[ESS/(n-2)]$.

- Can also define $R^2 = RSS/\sum_i (y_i - \overline{y})^2 = RSS/(RSS + ESS) =$ COEFFICIENT OF DETERMINATION. Thus $0 \le R^2 \le 1$.
  - If $R^2 \approx 1$, then ESS is small, so model $Y = b_1 + b_2 X$ is accurate, i.e. Y is heavily influenced by X.
  - If $R^2 \approx 0$, then RSS is small, so Y depends more on random effects than on $b_1 + b_2 X$, i.e. Y isn't influenced much by X.
  - THEOREM (Text Theorem 10.3.5): R^2 is the natural estimate of $[\mathrm{Corr}(X,Y)]^2 = [\mathrm{Cov}(X,Y)]^2 \,/\, \mathrm{Var}(X)\,\mathrm{Var}(Y)$. Indeed,

$$R^2 = \frac{(b_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2}{\sum_{i=1}^n (y_i - \overline{y})^2} = \frac{\left[\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})\right]^2}{\frac{1}{n-1} \sum_{i=1}^n (x_i - \overline{x})^2 \; \frac{1}{n-1} \sum_{i=1}^n (y_i - \overline{y})^2}$$

$$= \frac{\text{Estimate of } [\mathrm{Cov}(X,Y)]^2}{\text{Estimate of } \mathrm{Var}(X)\,\mathrm{Var}(Y)}\,.$$

—— **END WEDNESDAY 10** ——

[Reminder: Test #2 is 3–5 this Wednesday. Surnames A–Li in room CG 150, surnames Ll–Z in room CG 250. No aids allowed. Bring your T-Card!]

Previous Class:
* Reviewed Linear Regression, B_1, B_2, $B_1 + B_2 x$, S^2, F, RSS, ESS.
* Introduced "coefficient of determination", $R^2 = RSS/(RSS + ESS)$.

- NORMAL LINEAR REGRESSION:

- So far, we have generally assumed that $E[Y \mid X = x] = \beta_1 + \beta_2 x$, and (sometimes) that $\text{Var}[Y \mid X = x] = \sigma^2$.

- We now make a stronger assumption, that the conditional distribution of Y, given that $X = x$, is equal to $N(\beta_1 + \beta_2 x, \, \sigma^2)$. ["Normal Linear Regression", or "Linear Regression with Normal Errors".]

- In that case, we can determine many other distributions precisely [since linear combinations of normals are normal, etc.].

- Text Theorem 10.3.6: Under these assumptions, conditional on $X_1 = x_1, \ldots, X_n = x_n$,

$$B_1 \sim N\Big(\beta_1, \; \sigma^2 \big(\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\big)\Big) \, ;$$

$$B_2 \sim N\Big(\beta_2, \; \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\Big) \, ;$$

$$\frac{n-2}{\sigma^2} \, S^2 \sim \chi^2(n-2) \, ,$$

  with S^2 independent of (B_1, B_2).

  - So, $ESS/\sigma^2 \equiv \frac{n-2}{\sigma^2} \, S^2 \sim \chi^2(n-2)$.

  - By C.L.T., these distributions are <u>approximately</u> true for large n, even with other (non-normal) error distributions ...

- Then what about our F statistic?

  - Well, <u>if</u> $\beta_2 = 0$, then conditional on $X_1 = x_1, \ldots, X_n = x_n$,

$$(B_2) \sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 / \sigma^2} \;\sim\; N(0,1) \, ,$$

so
$$\frac{RSS}{\sigma^2} \;\equiv\; (B_2)^2 \sum_{i=1}^{n}(x_i - \overline{x})^2 \,/\, \sigma^2 \;\sim\; \chi^2(1)\,.$$

But $\frac{ESS}{\sigma^2} \equiv \frac{n-2}{\sigma^2}\, S^2 \sim \chi^2(n-2)$, so

$$F \;=\; \frac{(B_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2}{S^2} \;=\; \frac{[(B_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2 \,/\, \sigma^2] \,/\, (1)}{\frac{n-2}{\sigma^2}\, S^2 \,/\, (n-2)} \;\sim\; F(1,\, n-2)\,.$$

– But if $\beta_2 \neq 0$, then F should be <u>larger</u>.

– Hence, P-value for alternative hypothesis $\beta_2 \neq 0$, versus null hypothesis that $\beta_2 = 0$, is given by
$$P[W \geq \frac{(b_2)^2 \sum_{i=1}^n (x_i - \overline{x})^2}{s^2}]\,,$$
where $W \sim F(1, n-2)$. [Can compute from statistical package.]

– Note that mean of $F(a,b)$ is $b/(b-2)$, so mean of $F(1, n-2)$ is $(n-2)/(n-4) = 1 + 2/(n-4)$ [if $n > 4$], a little more than 1. [Makes sense, since we know that if $\beta_2 = 0$, then $F \approx 1$.]

– [Aside: Variance of $F(1, n-2)$ is $2(n-2)^2(n-3) \,/\, (n-4)^2(n-6) = O(1)$ as $n \to \infty$.]

- Example: Suppose observe pairs $(3,1)$, $(5,2)$, $(7,2)$, $(9,3)$. [DRAW GRAPH.]

  – Does Y increase with X (on average), or not? We want to test the null hypothesis that $\beta_2 = 0$ against the alternative hypothesis that $\beta_2 \neq 0$.

  – Compute (messy!) that $b_1 = 1/5$, $b_2 = 3/10$, $s^2 = 1/10$, and $F = 18$. [Exercise: Verify these!]

  – But expect that $F \sim F(1, n-2) = F(1,2)$.

  – Then P-value against null hypothesis ($\beta_2 = 0$) is given by $P[W \geq F] = P[W \geq 18]$, where $W \sim F(1,2)$.

  – We compute (from statistical package) that this P-value $\doteq 0.0513$. Thus, not <u>quite</u> 95% confident that observed increase wasn't just from chance. (But almost!)

  – [P-value would be smaller if n were larger.]

—— **END MONDAY 11** ——

[Held Test #2.]

Previous Class:

$*$ Normal Linear Regression:

—— Distribution of Y, conditional on $X = x$, is $N(\beta_1 + \beta_2 x,\, \sigma^2)$.

—— Then $B_2 \sim N\big(\beta_2,\ \sigma^2 \,/\, \sum_i (x_i - \overline{x})^2\big)$.

—— Also $B_1 \sim N\big(\beta_1,\ \sigma^2[(1/n) + (\overline{x})^2 \,/\, \sum_i (x_i - \overline{x})^2]\big)$.

—— Also $(n-2)S^2/\sigma^2 \sim \chi^2(n-2)$, indep. of B_1, B_2.

$*$ Then $ESS/\sigma^2 \sim \chi^2(n-2)$.

—— And $RSS/\sigma^2 \sim \chi^2(1)$ <u>if</u> $\beta_2 = 0$.

$*$ Thus $F \sim F(1, n-2)$ <u>if</u> $\beta_2 = 0$.

$*$ Then P-value for $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$ is given by $P[W \geq F]$, where F is observed value of F-statistic, and $W \sim F(1, n-2)$.

$*$ Example: Observe $(3, 1)$, $(5, 2)$, $(7, 2)$, $(9, 3)$.

—— Compute $b_1 = 1/5$, $b_2 = 3/10$, $s^2 = 1/10$, and $F = 18$.

—— Then P-value against $\beta_2 = 0$ is $P[W \geq 18] \doteq 0.0513$, where $W \sim F(1, n-2) = F(1, 2)$.

- Can also get <u>confidence intervals</u> for B_1 and B_2. [Here we focus on B_2.]

  - Since
  $$B_2 \sim N\big(\beta_2,\ \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\big)\,,$$
  therefore
  $$B_2 - \beta_2 \sim N\big(0,\ \frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2}\big)\,,$$
  so
  $$(B_2 - \beta_2)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \,/\, \sigma^2} \sim N(0, 1)\,.$$

  - But also $\frac{n-2}{\sigma^2} S^2 \sim \chi^2(n-2)$, independent of B_2. Hence,
  $$\frac{(B_2 - \beta_2)\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \,/\, \sigma^2}}{\sqrt{(\frac{n-2}{\sigma^2} S^2)/(n-2)}} \sim t(n-2)\,,$$

i.e.

$$(B_2 - \beta_2)\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2/S^2} \;\sim\; t(n-2)\,.$$

– So, if a_n is such that $P(-a_n < T_n < a_n) = 0.95$ whenever $T_n \sim t(n)$, then

$$P\Big[B_2 - a_{n-2}\sqrt{S^2\,/\,\sum_{i=1}^{n}(x_i - \overline{x})^2} < \beta_2 < B_2 + a_{n-2}\sqrt{S^2\,/\,\sum_{i=1}^{n}(x_i - \overline{x})^2}\Big] \doteq 0.95\,.$$

– i.e., $b_2 \pm a_{n-2}\sqrt{s^2\,/\,\sum_{i=1}^{n}(x_i - \overline{x})^2}$ is a 95% confidence interval for value of β_2.

- Above example continued:
  - Here $b_2 = 3/10$, $s^2 = 1/10$, and $\sum_{i=1}^{n}(x_i - \overline{x})^2 = 20$.
  - Also, $n = 4$, and if $T_2 \sim t(2)$, then $P[T_2 \le -4.3] \doteq 0.025$, so (by symmetry) $P[T_2 \ge +4.3] \doteq 0.025$, and $P[-4.3 < T_2 < +4.3] \doteq 1 - 0.025 - 0.025 = 0.95$, i.e. $a_2 \doteq 4.3$.
  - Hence, 95% confidence interval for β_2 is $(3/10) \pm 4.3\sqrt{(1/10)\,/\,20} = 0.3 \pm 4.3/\sqrt{200} \doteq 0.3 \pm 0.304 = (-0.004,\, 0.604)$.
  - This interval just <u>barely</u> contains 0. [Makes sense since $\beta_2 = 0$ is just <u>barely</u> possible at 95% confidence level.]

- B_1 is similar, since [Text Corollary 10.3.2]:

$$\frac{B_1 - \beta_1}{\sqrt{S^2\left(\frac{1}{n} + \frac{(\overline{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}} \;\sim\; t(n-2)\,.$$

- ONE CATEGORICAL PREDICTOR (10.4.1):
- Context:
  - In Categorical Response Models: both X and Y are categorical.
  - In Linear Regression: both X and Y are quantitative (i.e. numerical).

- Suppose now that Y is quantitative, but X takes values in one of a different catagories, $\{1, 2, \ldots, a\}$.
  - Example: Y = height, while X = gender (male or female).
  - For $i \in \{1, 2, \ldots, a\}$, let $\beta_i = E[Y \mid X = i]$. Want to estimate the β_i.
- Suppose for each $i \in \{1, 2, \ldots, a\}$, we observe n_i different values of Y corresponding to $X = i$, namely $y_{i1}, y_{i2}, \ldots, y_{in_i}$. Assume that $n_i \geq 1$ for all i.
  - Let $N = n_1 + n_2 + \ldots + n_a$ be total number of observations.
- How to estimate the β_i?
  - Use principle of least squares.
  - Here squared error is $SE = \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \beta_i)^2$.
  - Differentiable everywhere, goes to ∞ as any one $\beta_i \to \pm\infty$. So, SE is minimised at a critical point (if unique).
  - Critical point requires that $\frac{\partial}{\partial \beta_i} SE = 0$ for each i.
  - But $\frac{\partial}{\partial \beta_i} SE = -\sum_{j=1}^{n_i} 2(y_{ij} - \beta_i)$.
  - This equals 0 iff $n_i \beta_i = \sum_{j=1}^{n_i} y_{ij}$, i.e. $\beta_i = (1/n_i) \sum_{j=1}^{n_i} y_{ij} \equiv \overline{y}_i$, the average of the observations corresponding to $X = i$.
- Hence, estimate each β_i by the corresponding $\overline{y}_i$. [Makes sense.]
  - $E[\overline{Y}_i] = (1/n_i) \sum_{j=1}^{n_i} E[Y_{ij}] = (1/n_i) \sum_{j=1}^{n_i} \beta_i = \beta_i$. [Unbiased estimator.]
- What about variance? Suppose $\mathrm{Var}[Y \mid X = i] = \sigma^2$ for all i, but σ^2 is unknown. How to estimate?
  - Fact (Text Theorem 10.3.10): Unbiased estimate of σ^2 is given by

$$S^2 = \frac{1}{N - a} \sum_{i=1}^{a} \sum_{j=1}^{n_i} (y_{ij} - \overline{y}_i)^2 \, .$$

    [Like $\frac{1}{n-1}$ originally, and $\frac{1}{n-2}$ for linear regression. This time get to choose a values $(\overline{y}_1, \ldots, \overline{y}_a)$ based on data, which leads to the factor of $\frac{1}{N-a}$.]

- NORMAL ASSUMPTION: Assume now that the law of Y, given that $X = i$, is $N(\beta_i, \sigma^2)$.
  - Then $\overline{Y}_i \sim N(\beta_i,\, \sigma^2/n_i)$.
  - Also (Text Theorem 10.3.11), $(N-a)S^2/\sigma^2 \sim \chi^2(N-a)$, with S^2 independent of the $\overline{Y}_i$.
  - It follows that $\frac{\overline{Y}_i - \beta_i}{\sqrt{S^2/n_i}} \sim t(N-a)$.

—— **END MONDAY 12** ——

Previous Class:

∗ Normal Linear Regression:
—— Review.
—— Confidence Interval for β_2.
∗ One Categorical Predictor (10.4.1):
—— $X \in \{1, 2, \ldots, a\}$, $Y \in \mathbf{R}$.
—— Observe $n_i \geq 1$ observations with $X = i$; let $N = n_1 + \ldots + n_a$.
—— Least-squares estimate of $\beta_i \equiv E[Y \,|\, X = i]$ is $\overline{y}_i \equiv (1/n_i)\sum_{j=1}^{n_i} y_{ij}$. [Unbiased.]
—— Estimate $\sigma^2 \equiv \mathrm{Var}[Y \,|\, X = i]$ by $s^2 = (1/(N-a))\sum_i \sum_j (y_{ij} - \overline{y}_i)^2$. [Unbiased.]
∗ Normal assumption: Given $X = i$, $Y \sim N(\beta_i, \sigma^2)$.
—— Then $\overline{Y}_i \sim N(\beta_i,\, \sigma^2/n_i)$.
—— Also $(N-a)S^2/\sigma^2 \sim \chi^2(N-a)$ [assuming $N > a$].
—— Also $\frac{\overline{Y}_i - \beta_i}{\sqrt{S^2/n_i}} \sim t(N-a)$.

- CONFIDENCE INTERVALS:
  - Let a_n (again) be such that $P[-a_n < T_n < a_n] = 0.95$ whenever $T_n \sim t(n)$.
  - Then $P[-a_{N-a} < \frac{\overline{Y}_i - \beta_i}{\sqrt{S^2/n_i}} < a_{N-a}] = 0.95$.
  - Re-arranging, $P[\overline{Y}_i - a_{N-a}\sqrt{S^2/n_i} < \beta_i < \overline{Y}_i + a_{N-a}\sqrt{S^2/n_i}] = 0.95$.
  - i.e., 95% confidence interval for β_i is $\overline{y}_i \pm a_{N-a}\sqrt{s^2/n_i}$, where now $s^2 = \frac{1}{N-a}\sum_{i=1}^{a}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2$. [Interval depends on values of y_{kj} for $k \neq i$, too.]

- DIFFERENCES OF MEANS:

- What about differences $\beta_i - \beta_j$ $(j \neq i)$?
  - Well, $\overline{Y}_i \sim N(\beta_i,\, \sigma^2/n_i)$, and $\overline{Y}_j \sim N(\beta_j,\, \sigma^2/n_j)$, independent.
  - Therefore, $\overline{Y}_i - \overline{Y}_j \sim N(\beta_i - \beta_j,\ \sigma^2((1/n_i) + (1/n_j))$.
  - So,
$$\frac{(\overline{Y}_i - \overline{Y}_j) - (\beta_i - \beta_j)}{\sqrt{\sigma^2((1/n_i) + (1/n_j))}} \ \sim\ N(0,1)\,.$$
  - Here σ^2 is unknown (as usual). But we know that $(N-a)S^2/\sigma^2 \sim \chi^2(N-a)$.
  - So, $T \sim t(N-a)$, where
$$T \ =\ \frac{\frac{(\overline{Y}_i - \overline{Y}_j) - (\beta_i - \beta_j)}{\sqrt{\sigma^2((1/n_i)+(1/n_j))}}}{\sqrt{((N-a)S^2/\sigma^2)/(N-a)}} \ =\ \frac{(\overline{Y}_i - \overline{Y}_j) - (\beta_i - \beta_j)}{\sqrt{S^2((1/n_i) + (1/n_j))}}\,.$$
  - Then $P[-a_{N-a} < T < a_{N-a}] = 0.95$.
  - Hence,
$$P[-a_{N-a} < \frac{(\overline{Y}_i - \overline{Y}_j) - (\beta_i - \beta_j)}{\sqrt{S^2((1/n_i) + (1/n_j))}} < a_{N-a}] = 0.95\,.$$
  - Re-arranging, $P[(\overline{Y}_i - \overline{Y}_j) - a_{N-a}\sqrt{S^2((1/n_i) + (1/n_j))} < \beta_i - \beta_j < (\overline{Y}_i - \overline{Y}_j) + a_{N-a}\sqrt{S^2((1/n_i) + (1/n_j))}] = 0.95$.
  - Thus, 95% confidence interval for $\beta_i - \beta_j$ is $(\overline{y}_i - \overline{y}_j) \pm a_{N-a}\sqrt{s^2((1/n_i) + (1/n_j))}$.
- EXAMPLE:
  - Suppose measuring IQs of students at U of T and at York. U of T students: 130, 150, 140, 150, 170, 160. York students: 130, 140, 135.
  - Then $\overline{y}_1 = (130 + 150 + 140 + 150 + 170 + 160)/6 = 150$. And $\overline{y}_2 = (130 + 140 + 135)/3 = 135$. Also $n_1 = 6$ and $n_2 = 3$, and $N = 6 + 3 = 9$, and $a = 2$.
  - Then $s^2 = \frac{1}{N-a}\sum_{i=1}^{2}\sum_{j=1}^{n_i}(y_{ij} - \overline{y}_i)^2 = \frac{1}{9-2}\Big((130-150)^2 + (150-150)^2 + (140-150)^2 + (150-150)^2 + (170-150)^2 + (160-150)^2 + (130-135)^2 + (140-135)^2 + (135-135)^2\Big) = \frac{1}{7}(400 + 0 + 100 + 0 + 400 + 100 + 25 + 25 + 0) = 1050/7$.
  - Also, if $T_7 \sim t(7)$, then $P[-2.36 < T_7 < 2.36] \doteq 0.95$.

- Thus, 95% confidence interval for mean of U of T IQs (i.e., β_1) is given by $\overline{y}_1 \pm 2.36\sqrt{s^2/n_1} = 150 \pm 2.36\sqrt{(1050/7)/6} \doteq (138.2, 161.8)$.
- And, 95% confidence interval for mean of York IQs (i.e., β_2) is given by $\overline{y}_2 \pm 2.36\sqrt{s^2/n_2} = 135 \pm 2.36\sqrt{(1050/7)/3} \doteq (123.2, 146.8)$.
- Some overlap in these intervals. What about <u>difference</u>?
- Here 95% confidence interval for difference $\beta_1 - \beta_2$ is given by $(\overline{y}_1 - \overline{y}_2) \pm 2.36\sqrt{s^2((1/n_1) + (1/n_2))} = (150-135) \pm 2.36\sqrt{(1050/7)[(1/6) + (1/3)]} \doteq (-5.4, 35.4)$.
- So, <u>probably</u> $\beta_1 > \beta_2$, i.e. average IQ at U of T is larger than average IQ at York, but we're not quite 95% sure that it is.

- Final Exam is Monday, May 3, 9:00 a.m. – 12:00 noon, in University College, East Hall (surnames A-Li) and West Hall (surnames Ll-Z).

- FINAL COMMENT: Statistics courses for next year.
  - STA 302, STA 322, STA 322: More about <u>applied statistics</u> techniques. [regression analysis / sample surveys / experimental design]
  - STA 352: More about the mathematical <u>theory</u> of statistical inference.
  - STA 347: More about <u>probability</u> theory (expand on STA 257).
