4. $\delta(X_1, \ldots, X_n) = \text{median}(X_1, \ldots, X_n)$, estimating $\theta =$ median
   a. Median for sample of size $2n + 1$ has a density $\frac{(2n+1)!}{n!n!} F(x)^n (1 - F(x))^n f(x)$
   b. Consider exponential model with CDF $F(x) = 1 - \exp(-x)$.
   c. Value of population median is $\log(2) = 0.693147$
   d. Integral for $\mathrm{E}[\text{median}]$ can be done in closed form.
   e. Table 1 has results

Table 1: True Expectation for the Exponential Median

Sample Size	Expectation of Median
3	0.833333
5	0.783333
7	0.759524
9	0.745635
11	0.736544
∞	0.693147

E. General Rule about Unibiased Estimators

1. For identically distributed observations, mean is unbiased, without regard to the distribution.
   a. This does not require independence.
2. Convex combination of unbiased estimators is unbiased estimator

a. Convex combination is the sum of items combined times nonnegative constants, with constants summing to 1.

3. Hence sample average is always an unbiased estmator of the expectation of each observation, if expectation exists.
4. If $\delta(X)$ is an unbiased estimator of θ, then $a + b\delta(X)$ is an unbiased estimator of $a + b\theta$
5. If $\delta(X)$ is an unbiased estimator of θ, and f is a transformation not of the form $x \mapsto a + bx$, then $f(\delta(X))$ is generally a biased estimator of $f(\theta)$.

WMS: 8.5

F. Confidence Intervals

1. Example: Paleontology.
   a. Goal: estimate how long ago a certain species of animal first walked or crawled the earth.
   b. You assume
      i. species population has been constant since its advent θ years ago,
      ii. the probability of finding any one of these animals is the same regardless of its age.

c. Completely unreasonable assumptions imply that the age X of a given sample $\sim \mathcal{U}(0, \theta)$.

d. You date n specimens.

i. $\max X_j$ is biased, but $(\max X_j)(n+1)/n$ is unbiased.

2. Confidence Interval goal is to give a range containing true parameter.

a. Extreme Answers:

i. Probability of hitting the true value on the head is zero

ii. In order to get a range of possible values that will always include the true value, we'd have to take the whole parameter domain.

b. Compromise solution is to look for bounds $\theta_L(X_1, \cdots, X_n)$ and $\theta_U(X_1, \cdots, X_n)$ such that $\theta_L(X_1, \cdots, X_n)$ will fall below the parameter and that $\theta_U(X_1, \cdots, X_n)$ will fall above the parameter with a certain probability.

c. If such an $\theta_L(X_1, \cdots, X_n)$ and $\theta_U(X_1, \cdots, X_n)$ exist $(\theta_L(X_1, \cdots, X_n), \theta_U(X_1, \cdots, X_n))$ is called a $confidence\ interval$ (c.i.)

i. In symbols, $\mathrm{P}\left[\theta_L \leq \theta \leq \theta_U\right] \geq 1 - \alpha$.

- $1 - \alpha$ called confidencce level.

- Most often, $1 - \alpha = .95 = 95\%$.

3. Strategy: Manipulate a probability statement about the parameter of interest and a statistic that the interval end points are likely to be a function of.

   a. Graphical strategy:

      i. Choose $t_1(\theta)$ and $t_2(\theta)$ such that $\forall \theta$, $\mathrm{P}\left[T \geq t_1(\theta)\right] \geq .975$ and $\mathrm{P}\left[T \leq t_2(\theta)\right] \geq .975$.

      ii. Here set $t_1(\theta) = F^{-1}(.025; \theta)$ and $t_2(\theta) = F^{-1}(.975; \theta)$.

      iii. For each potential value of $\hat{\theta}$,

      - draw a horizontal line between the curves.
      - This will be the c.i. as a function of $\hat{\theta}$.

      iv. If $t_2(\theta)$ and $t_1(\theta)$ are increasing in θ the vertical line above θ from $t_2(\theta)$ to $t_1(\theta)$ is the c.i..

      v. How often will this cover θ?

      - c.i. covers θ if and only if the vertical and horizontal lines cross,
      - if and only if $\hat{\theta}$ lies between t_2 and t_1,
      - happens $1 - \alpha$ of the time.

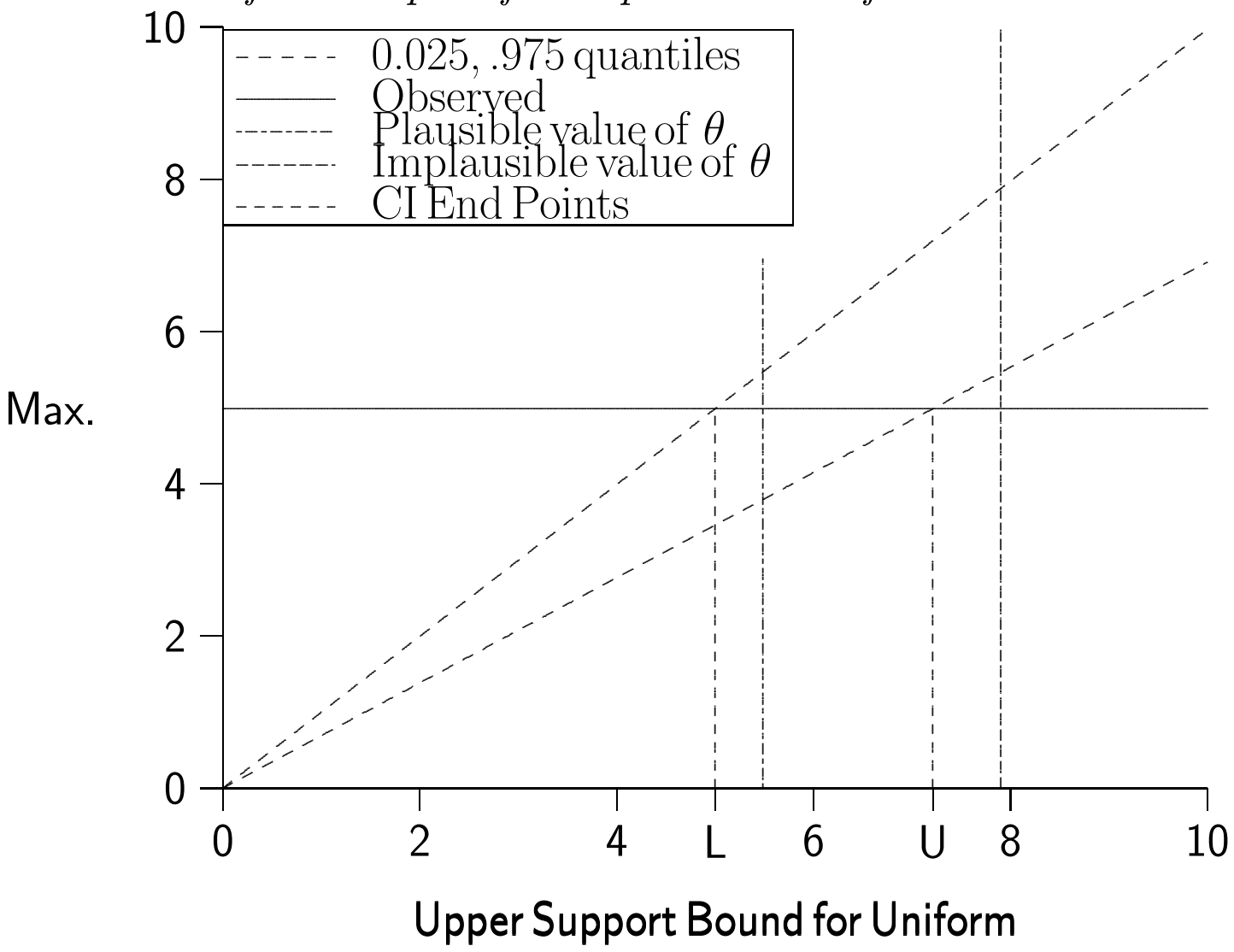
   b. Fig. 2 shows graphical construction for continuous variable

c. Fig. 3 shows graphical construction for discrete variable

i. CDF has flat parts

ii. CI runs to ends of flat parts in such a way as to make the intervals wider.

Fig. 2: Confidence Interval Construction for Sample of Independent Uniforms



d. Algebraic strategy:

i. Choose S a function of T and θ such that the distribution of S does not depend on θ.

ii. we say S is $pivotal$.

iii. Let $F_S(s)$ be the c.d.f. of S

iv. Solve $F_S(s_{.975}) = .975$ and $F(s_{.025}) = .025$.

v. Construct intervals for the pivotal quantity, and solve for θ.

4. Example: n $\mathcal{U}[0, \theta]$ variables,

a. Let $S = T/\theta$

b. cumulative distribution function of T is $F(t; \theta) = t^n/\theta^n$ if $t \leq \theta$.

c. $s^n_{.95} = .95$ or $s_{.95} = \sqrt[n]{.95}$ and $s^n_{.05} = .05$ or $s_{.05} = \sqrt[n]{.05}$.

d. Hence $\mathrm{P}\left[T < \theta \sqrt[n]{.05}\right] = \mathrm{P}\left[T > \theta \sqrt[n]{.95}\right] = .05$.

e. Hence $\mathrm{P}\left[\theta \sqrt[n]{.05} \leq T \leq \theta \sqrt[n]{.95}\right] = .90$.

f. Hence $\mathrm{P}\left[T \, {}^{-n}\!\sqrt{.05} \leq \theta \leq T \, {}^{-n}\!\sqrt{.95}\right] = .90$.

g. Works because $S = T/\theta$ has a distń func. ind. of what we were trying to estimate or other unknown parameters;

WMS: 8.6

G. Confidence Intervals Using the Normal Distribution

1. Normal Distribution with known variance. $T = \bar{X}$;

$$F_T(t;\mu) = \Phi((t-\mu)\sqrt{n}\sigma^{-1}).$$

a. Hence $(T-\mu)\sqrt{n}\sigma^{-1}$ is pivotal.

b. Confidence interval is $T \pm \sigma z_{\alpha/2}/\sqrt{n}$

i. Here z_β is the number with probability β above it in the normal table.

$$\begin{aligned} 1-\alpha &= \mathrm{P}\left[-z_{\alpha/2} \le (T-\mu)\sqrt{n}\sigma^{-1} \le z_{\alpha/2}\right] \\ &= \mathrm{P}\left[-\sigma z_{\alpha/2}/\sqrt{n} \le T-\mu \le \sigma z_{\alpha/2}/\sqrt{n}\right] \\ &= \mathrm{P}\left[\sigma z_{\alpha/2}/\sqrt{n} \ge \mu - T \ge -\sigma z_{\alpha/2}/\sqrt{n}\right] \\ &= \mathrm{P}\left[T+\sigma z_{\alpha/2}/\sqrt{n} \ge \mu \ge T-\sigma z_{\alpha/2}/\sqrt{n}\right] \end{aligned}$$

2. Formula generally works if σ must be replaced by estimate.

a. Ex. $X_1, \ldots, X_n$ iid, unknown variance σ^2.

i. Estimate σ^2 by $S^2 = \Sigma_{i=1}^n (X_i - \bar{X})^2/(n-1)$.

ii. Hence approximate CI is $\bar{X} \pm z_{\alpha/2} S/\sqrt{n}$.

b. Ex. $X \sim \mathcal{B}\mathrm{in}(\theta, n)$, $\hat{\theta} = X/n$.

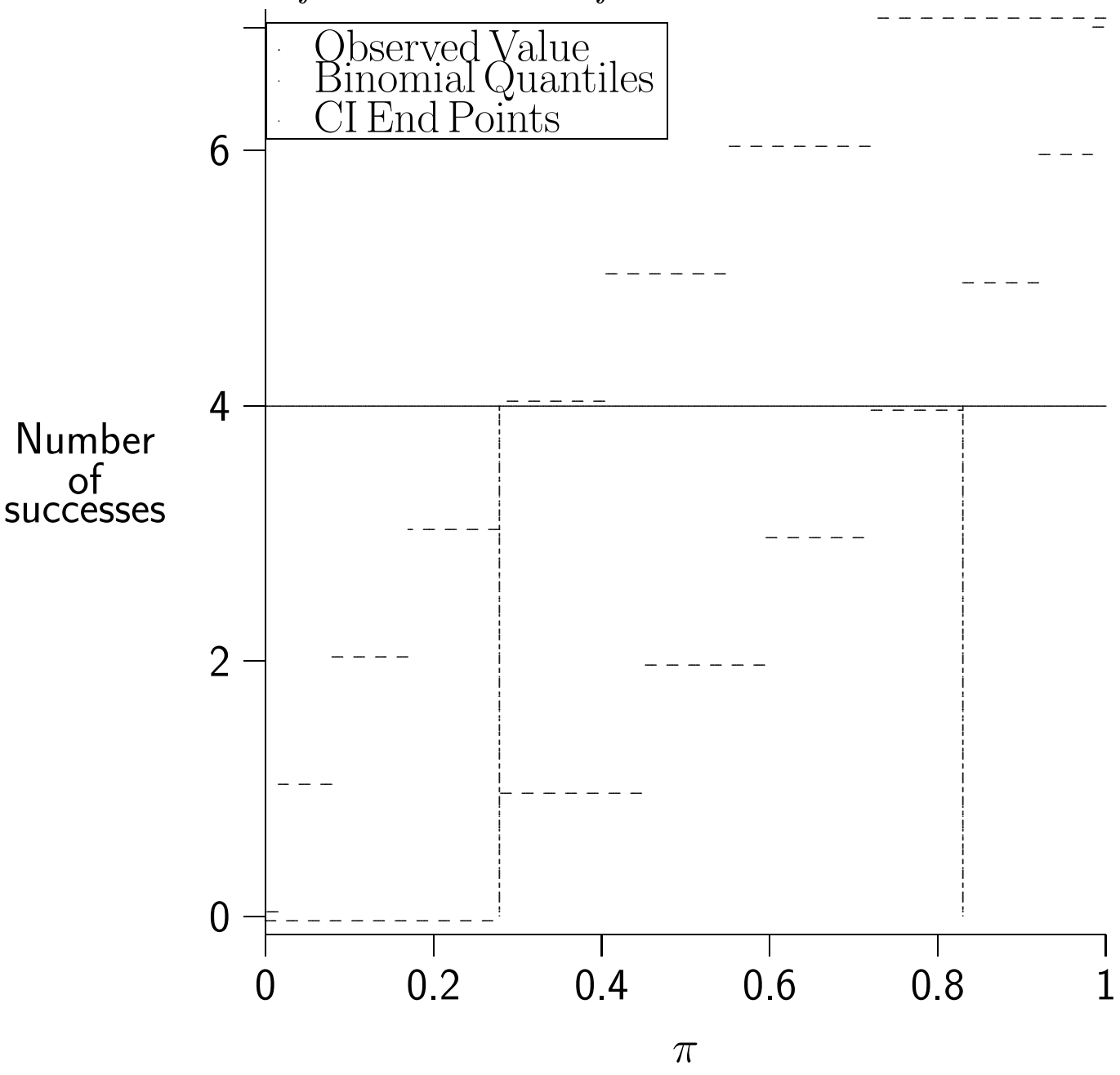
i. $(\hat{\theta}-\theta)/\sqrt{\theta(1-\theta)/n} \sim \mathcal{N}(0,1)$.

ii. $(\hat{\theta}-\theta)/\sqrt{\hat{\theta}(1-\hat{\theta})/n} \sim \mathcal{N}(0,1)$.

iii. Using the above rule, 95% CI for θ is $(\hat{\theta} - 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}, \hat{\theta} + 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n})$

iv. Alternatively, use as interval $\{\theta | (\hat{\theta}-\theta)^2/(\theta(1-\theta)/n) \le$

Fig. 3: Graphical Construction of Binomial Confidence Interval



$z^2_{\alpha/2}\}$.

- Expressible as (θ_L, θ_U), where endpoints are solution to quadratic equation.

- Very close to plug-in solution.

WMS: 8.7

H. Sample size:

1. Suppose you want to estimate parameter to within a certain accuracy e
   a. called $margin\,of\,error$.
2. As measured by ci of level $1-\alpha$.
3. Suppose you have pre-knowledge of the standard deviation.
4. Then $\sigma z_{\alpha/2}/\sqrt{n} \leq e$
5. Then $\sigma z_{\alpha/2}/e \leq \sqrt{n}$
6. Then $n \geq \sigma^2 z_{\alpha/2}^2/e^2$
   a. Ex., to estimate binomial proportion (ex. poll result) to 2%,
      i. $\sigma^2 = \theta(1-\theta) \leq .25$
      ii. Can get by with $n = .25(1.96)^2/.02^2 \approx 2500$.

WMS: 8.8-8.9

I. Common Applications

1. Above we saw one-sample binomial and means confidence intervals
2. Two-sample mean difference

a. Assume

i. $X_1, \ldots, X_m$ same expectation and finite variance

ii. $Y_1, \ldots, Y_n$ same expectation and finite variance

iii. All independent

b. Estimate $\theta = \mathrm{E}[Y_i] - \mathrm{E}[X_i]$

i. Case with common variance:

- Pivotal quantity $S = (\bar{Y} - \bar{X} - \theta)/(\sigma\sqrt{1/m + 1/n})$ if common variance were known to be σ.
- Pivotal quantity $S = (\bar{Y} - \bar{X} - \theta)/(S_p\sqrt{1/m + 1/n})$
- $S_p = \sqrt{(\Sigma_{i=1}^{m}(X_i - \bar{X}) + \Sigma_{i=1}^{n}(Y_i - \bar{Y}))/(m + n - 2)}$: pooled standard deviation.
- Pivot has distribution approximately $\mathcal{N}(0, 1)$
  - ▷ More closely, t_{m+n-2}.