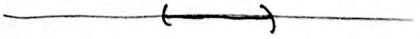
CHAPTER 7: INTERVAL ESTIMATION

- Using MLE's we obtained point estimators with important properties.

- However, for a given sample, we might want more, say, an interval of values instead of a single point-value.

Point Estimator: realization of a well-behaved estimator r.v.

Confidence Region: a subset of $\mathbb{R}^n$, where the estimator captures the parameter with a desired probability

Confidence Interval:  a confidence region that has the form of an interval.

- Confidence regions/intervals are only possible to construct if we know the distributional properties of the estimator.

- This will be particularly possible if the estimator is $\bar{X}$, due to CLT.

7.1 Confidence Intervals for Means

- We'll proceed from the case where we know the most to cases where we know less & less to see how we can construct the confidence intervals.

Case 1: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$. $\hat{\mu} = \overline{X} \sim N(\mu, \frac{\sigma^2}{n})$ (MLE)

μ unknown

σ^2 known

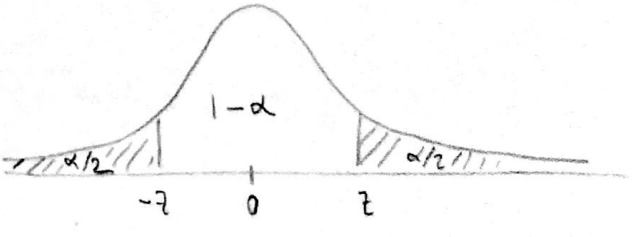
$$\Rightarrow$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1).$$

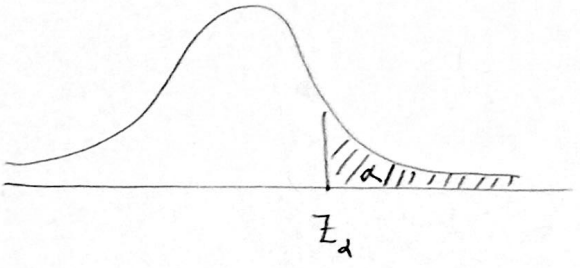
- Our goal is to find an interval I s.t. $P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \in I\right) \geq 1-\alpha$, where this $1-\alpha$ is a large probability.

- For practical purposes we might want I to be centered @ 0 (and be symmetric) as well as we'll be focusing on the smallest I that satisfies this relationship, i.e., $P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \in I\right) = 1-\alpha$.

- Result:



- In Section 3.3 (p113), we defined z_α to be the value that gives the right-tail probability of $\alpha \in [0,1]$:



as $\alpha \uparrow$ $z_\alpha \downarrow$

- In other words, for $Z \sim N(0,1)$, $P(Z \geq z_\alpha) = \alpha$

iff

$\Phi(z_\alpha) = 1 - \alpha$

iff

$\Phi(-z_\alpha) = \alpha$

iff

$P(Z \geq -z_\alpha) = 1 - \alpha$

- A few values:

$z_{0.0125} = 2.24$

$z_{0.025} = 1.96$

$z_{0.05} = 1.645$

(see Table Vb, p503)

- Finally we have the complete picture:

$$P\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \in [-z_{\alpha/2}, z_{\alpha/2}]\right) = P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1-\alpha$$

- Remember however, we want to "catch" μ. So, modify:

$$P\left(-z_{\alpha/2} \leq \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = P\left(\bar{X} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right)$$

$$= 1-\alpha.$$

- Now, we have the random interval: $\left[\bar{X} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}}\right]$

which includes μ with $1-\alpha$ probability.

- We'll call the realization of this random interval a

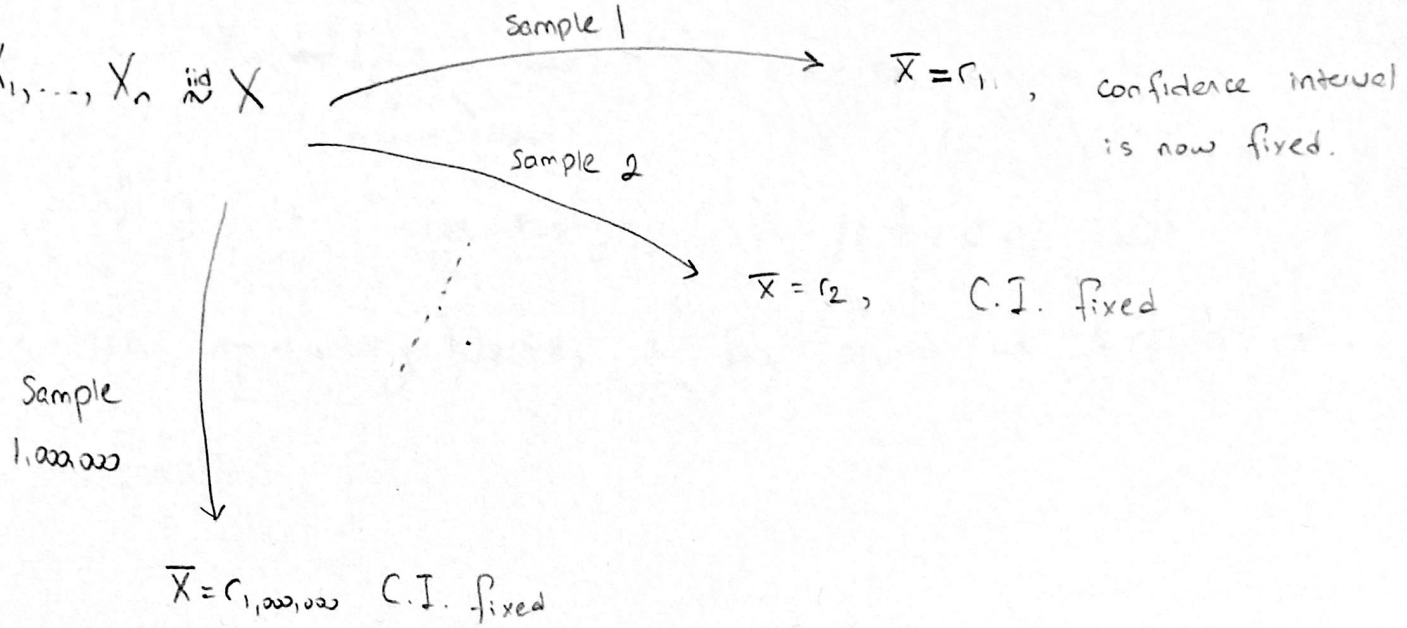
<u>$100\cdot(1-\alpha)\%$ confidence interval.</u>

- Note: once $\bar{X}$ is realized, there is no more randomness.

$$\mu \in \mathbb{R}, \quad \bar{x} - z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \in \mathbb{R}, \quad \bar{x} + z_{\alpha/2}\cdot\frac{\sigma}{\sqrt{n}} \in \mathbb{R}$$

So, μ is either in the interval or not. We might only argue:

$$P(\mu \in I) \text{ is } 0 \text{ or } 1.$$



- 95% C.I. means, before realization there was 95% chance of including μ. But once we fix the C.I. it either has μ or not.

- Roughly 95% of the above 1,000,000 C.I.'s will include μ.

HW Study Ex 7.1-2. $\frac{1-0.95}{2} = \frac{0.05}{2} = 0.025$

Ex: X = future lifetime of a 60-watt light bulb, $X \sim N(\mu, \underbrace{1296}_{\sigma^2})$.

$n = 27$, $\bar{X} = 1478$ $\Rightarrow$ a 95% C.I:

$$\left[\bar{X} - z_{0.025}\cdot\frac{\sigma}{\sqrt{n}},\ \bar{X} + z_{0.025}\cdot\frac{\sigma}{\sqrt{n}}\right] = \left[1478 - 1.96\cdot\frac{36}{\sqrt{27}},\ 1478 + 1.96\cdot\frac{36}{\sqrt{27}}\right]$$

$z_{0.025}$ ↓ negative one

$$= [1464.42,\ 1491.58]$$

Case 2: $X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$ but $\bar{X}$ is not exactly Normal.

(μ: unknown, σ^2: known)

If n is "large enough", $(\bar{X} - \mu)/\sigma/\sqrt{n}$ will be approximately Normal, so an approximate C.I. is still feasible.

(i) If X_i are unimodal, symmetric, continuous, $n \geq 5$ provides quite a good approximation.

(ii) If X_i are badly skewed or discrete, you want a larger sample.

- We'll assume $n \geq 30$ to be acceptable.

Case 3: $X_1, \ldots, X_n \overset{iid}{\sim} (\mu, \sigma^2)$. We cannot work with $\dfrac{\bar{X} - \mu}{\sigma/\sqrt{n}}$ anymore.

(μ: unknown, σ^2: unknown)

Instead, we'll use S^2 substituting σ^2. Recall $S^2 = \dfrac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$

- In this case we have two approximations: $\bar{X} \approx$ Normal

  $S^2 \approx \sigma^2$

- $n \geq 30$ will be good enough unless X_i's are badly skewed, in which case we must have $n \geq 50$ or higher.

- Difference from the previous set up:

$$\bar{X} \mp z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \quad \rightsquigarrow \quad \bar{X} \mp z_{\alpha/2} \cdot \frac{S}{\sqrt{n}},$$

where S, similar to $\bar{X}$, is calculated from the sample values.

Case 4: $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu, \sigma^2)$ and n is small.

(μ: unknown, σ^2: unknown)

$$T = h(X_1, \ldots, X_n) = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{t-distribution with } n-1 \text{ degrees of freedom},$$ see Section 7.5, p205.

- Similar to how $z_{\alpha/2}$ works, we'll choose $t_{\alpha/2} \in \mathbb{R}$ s.t.

$$P\left(T \geq t_{\alpha/2,(n-1)}\right) = \alpha/2.$$

For these $t_{\alpha/2}$, see Table VI on p 504.

- Then the $100(1-\alpha)\%$ C.I. for μ will be

$$\left[\bar{X} - t_{\alpha/2,(n-1)} \cdot \frac{S}{\sqrt{n}},\ \bar{X} + t_{\alpha/2,(n-1)} \cdot \frac{S}{\sqrt{n}}\right]$$

Case 5: $X_1, \ldots, X_n \sim (\mu, \sigma^2)$ are not Normal, are badly skewed and n is small

↓

it would be better to use the nonparametric approaches we'll see in Section 7.5.

HW As another special case of confidence regions, read on the one-sided confidence intervals, on p 312.

HW 1-5, 7-13

7.2 Confidence Intervals for the Difference of Two Means

- In the previous section, we constructed C.I.'s that aim to "catch" the unknown parameter, μ.
- In a sense, we are establishing the relationship bw. $\bar{X}$ & the real number μ.
- What if, instead, we want to compare two different sample means, say $\bar{X}$ v.s. $\bar{Y}$? What kind of r.v. would their difference, $\bar{X} - \bar{Y}$ would be? Can we construct C.I.'s to understand what the difference $\mu_x - \mu_y$ is ? If we can, we may draw conclusions about how different μ_x v.s. μ_y are. This has a wide variety of applications.

Case 1: $X_1, \dots, X_{n_x} \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$ and $Y_1, \dots, Y_{n_y} \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$

(μ_x, μ_y: unknown; σ_x^2, σ_y^2: known)

Here $X_i \perp\!\!\!\perp Y_j$ for all i, j.

$$\Rightarrow \bar{X} \perp\!\!\!\perp \bar{Y}$$

$$\Rightarrow \bar{X} - \bar{Y} \sim N\left(\mu_x - \mu_y, \frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}\right)$$

- This means $$P\left(-z_{\alpha/2} \leq \frac{(\bar{X}-\bar{Y})-(\mu_X-\mu_Y)}{\sqrt{\sigma_X^2/n_X + \sigma_Y^2/n_Y}} \leq z_{\alpha/2}\right) = 1-\alpha,$$

and so, the $100(1-\alpha)\%$ C.I. for $\mu_X - \mu_Y$ is

$$\left[(\bar{x}-\bar{y}) - z_{\alpha/2}\cdot\sigma_w \;,\; (\bar{x}-\bar{y}) + z_{\alpha/2}\cdot\sigma_w\right],$$

where $\sigma_w^2 = \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}$ is the variance of $\bar{X}-\bar{Y}$.

Ex: Say $\left\{\begin{matrix} n_X = 15 \\ \bar{x} = 70.1 \end{matrix}\right\}$ and $\left\{\begin{matrix} n_Y = 8 \\ \bar{y} = 75.3 \end{matrix}\right\}$ and $(1-\alpha) = 0.9$

& &

$\sigma_X^2 = 60$ $\sigma_Y^2 = 40$

$1-\alpha/2 = 0.95 \Rightarrow z_{\alpha/2} = 1.645$

$\sigma_w^2 = \frac{60}{15} + \frac{40}{8} = 9$

$\Big\}\; z_{\alpha/2}\cdot\sigma_w = 4.935$

$\bar{x}-\bar{y} = -5.2$ $\therefore$ 90% C.I. : $\left[-5.2-4.935, -5.2+4.935\right] = \left[-10.135, -0.265\right]$

- This gives strong evidence to suggest: $\overline{x} < \overline{y}$

Case 2: $X_1, \dots, X_{n_X} \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$, $Y_1, \dots, Y_{n_Y} \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$

(μ_X, σ_X^2: unknown; μ_Y, σ_Y^2: unknown)

- The sample sizes, n_X and n_Y are large.

- In this case, we may use S_X^2 and S_Y^2 in the place of σ_X^2, σ_Y^2.

So the approximate $100(1-\alpha)\%$ C.I. for $\mu_X - \mu_Y$:

$$\overline{x} - \overline{y} \mp z_{\alpha/2} \cdot \sqrt{\frac{S_X^2}{n_X} + \frac{S_Y^2}{n_Y}}$$

Case 3: $X_1, \dots, X_n \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$, $Y_1, \dots, Y_n \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$,

(μ_X, σ_X^2: unknown; μ_Y, σ_Y^2: unknown)

- The sample sizes, n_X and n_Y are small. We cannot hope to use S_X^2, S_Y^2 instead of σ_X^2, σ_Y^2 accurately in this case.

- There's still wiggle room if we know $\sigma_X^2 = \sigma_Y^2 =: \sigma^2$, despite not knowing the actual value σ^2.

- In this case

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{\frac{\sigma^2}{n_X} + \frac{\sigma^2}{n_Y}}} \sim N(0,1)$$

- Further, consider

$$U = \frac{(n_X - 1) \cdot S_X^2}{\sigma^2} + \frac{(n_Y - 1) \cdot S_Y^2}{\sigma^2} \sim \chi^2(n_X + n_Y - 2)$$

$\downarrow$ $\chi^2(n_X - 1)$ $\qquad$ $\downarrow$ $\chi^2(n_Y - 1)$

- By Theorem 5.5-2, Z and U are independent. Why?

$$Z = g(\bar{X}, \bar{Y}), \quad U = h(S_X^2, S_Y^2)$$

- Define $T = \dfrac{Z}{\sqrt{U/(n_X + n_Y - 2)}} \sim$ t distribution with $n_X + n_Y - 2$ d.o.f., by Theorem 5.5-3.

$$T = \frac{Z}{\sqrt{U/(n_x+n_y-2)}} = \frac{\dfrac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\sqrt{\sigma^2/n_x+\sigma^2/n_y}}}{\sqrt{\left(\dfrac{(n_x-1)S_x^2}{\sigma^2}+\dfrac{(n_y-1)\cdot S_y^2}{\sigma^2}\right)\Big/(n_x+n_y-2)}}$$

$$= \frac{\bar{X}-\bar{Y}-(\mu_x-\mu_y)}{\left[\dfrac{(n_x-1)\cdot S_x^2+(n_y-1)\cdot S_y^2}{n_x+n_y-2}\right]\cdot\left[\dfrac{1}{n_x}+\dfrac{1}{n_y}\right]} \sim t \text{ distribution with } n_x+n_y-2 \text{ d.o.f.}$$

- The key point here is eliminating the dependence of σ^2's, while knowing the distribution.
- So, we have

$$P\left(-t_{\alpha/2}\cdot(n_x+n_y-2) \leq T \leq t_{\alpha/2}\cdot(n_x+n_y-2)\right) = 1-\alpha$$

- Solving for $\mu_x-\mu_y$, we obtain the $100(1-\alpha)\%$ C.I. :

$$\bar{x}-\bar{y} \mp t_{\alpha/2}\cdot(n_x+n_y-2)\cdot S_p\cdot\sqrt{\frac{1}{n_x}+\frac{1}{n_y}}, \text{ where}$$

S_p is the realized value of $S_p = \sqrt{\dfrac{(n_x-1)\cdot S_x^2+(n_y-1)\cdot S_y^2}{n_x+n_y-1}}$.

- Case 3 may be extended from $\frac{\sigma_X^2}{\sigma_Y^2} = 1$ $\rightsquigarrow$ $\frac{\sigma_X^2}{\sigma_Y^2} = d > 0$

See Exc. 7.2-8.

Case 4: Same setup, but we don't know if $\frac{\sigma_X^2}{\sigma_Y^2} = d$ for some $d > 0$. If n_X, n_Y are large, this is Case 2. But this does not apply if the sample sizes are not large enough. (unlike Case 3)

One particular issue in this case if for n_X or n_Y small, S_X^2 or S_Y^2 ends up being large. Aspin (1949) provided the t distribution that works to prevent this, with r degrees of freedom, where

$$\frac{1}{r} = \frac{c^2}{n_X - 1} + \frac{(1-c^2)}{n_Y - 1} \quad \text{and} \quad c = \frac{S_X^2/n_X}{S_X^2/n_X + S_Y^2/n_Y}$$

See Eqn 7.2-1 for another equivalent expression.

- In case $r \notin \mathbb{Z}$, use $\lfloor r \rfloor$. $W = \frac{\bar{X} - \bar{Y} - (\mu_X - \mu_Y)}{\sqrt{S_X^2/n_X + S_Y^2/n_Y}} \approx t(\lfloor r \rfloor)$

- The cases so far do not include $\{X_i\}_{i=1}^{n_x}$ and $\{Y_j\}_{j=1}^{n_y}$ that may be dependent. In this case the above arguments/methods fail to hold true as the r.v. we worked with fail to be (approximately) t or χ^2.

Case 5: Let $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ be n pairs of dependent measurements. $X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_x, \sigma_x^2)$, $Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu_y, \sigma_y^2)$. Define $D_i = X_i - Y_i$, for $i = 1, 2, \ldots, n$. Then $D_1, \ldots, D_n$ may be considered as a random sample from $N(\mu_D, \sigma_D^2)$, where

$$\mu_D = \mu_x - \mu_y \quad \text{and} \quad \sigma_D^2 = \sigma_x^2 + \sigma_y^2.$$

$$T = \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}}$$

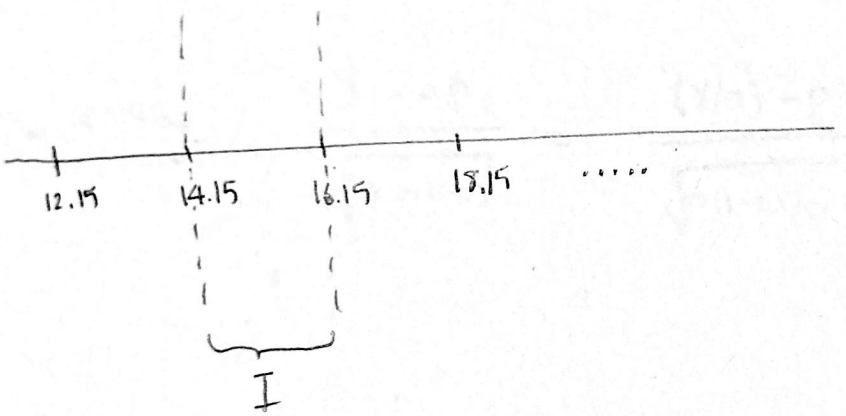
has a t-distribution with $n-1$ d.o.f. Thus the $100(1-\alpha)\%$ C.I. for $\mu_D = \mu_x - \mu_y$ is:

$$\bar{d} \mp t_{\alpha/2} \cdot (n-1) \cdot \frac{S_d}{\sqrt{n}}$$

HW 1-5, 10-12

7.3 – CONFIDENCE INTERVALS FOR PROPORTIONS

- Consider the class intervals in the midterm, specifically, the second one:



- Take a sample of size $n = 50$. Say for a single realization of the random sample $X_1, \ldots, X_{50}$, the probability that it'll fall into $I = (14.15, 16.15)$ is p.

- Then the number of sample values that will fall into I, has a binomial distribution:

$$Y \sim Bin(n, p).$$

- The distribution of the relative frequency of I: $\frac{Y}{n}$

- $\mathbb{E}\left[\frac{Y}{n}\right] = \frac{1}{n} \cdot \mathbb{E}[Y] = \frac{1}{n} \cdot np = p \rightsquigarrow$ unbiased estimator of p.

- $\frac{Y}{n}$ being an unbiased point estimator, one wonders if we can construct confidence intervals.

- We begin by recalling $\dfrac{Y - n.p}{\sqrt{n.p.(1-p)}} = \dfrac{(Y/n) - p}{\sqrt{p(1-p)/n}} \approx N(0,1)$, if n is large enough.

- By Table V in the appendix we may find $z_{\alpha/2}$ s.t.

$$P\left(-z_{\alpha/2} \leq \frac{(Y/n) - p}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}\right) \approx 1 - \alpha$$

$\rightsquigarrow$ Solve for p

$$P\left(\frac{Y}{n} - z_{\alpha/2} \cdot \sqrt{\frac{p.(1-p)}{n}} \leq p \leq \frac{Y}{n} + z_{\alpha/2} \cdot \sqrt{\frac{p.(1-p)}{n}}\right) \approx 1 - \alpha$$

- The main problem: p is not only in the center, but also in the endpoints.

① One way to fix it: if n is large enough

$$p \longleftrightarrow \hat{p} = \frac{Y}{n} = \frac{y}{n}$$

(y: realized)

$$100(1-\alpha)\%\ \text{C.I.}: \quad \hat{p} \mp z_{\alpha/2} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

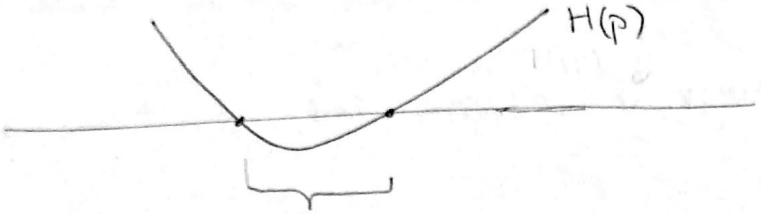
② Alternatively,

$$\frac{|Y/n - p|}{\sqrt{p(1-p)/n}} \leq z_{\alpha/2}$$

iff

$$\underbrace{\left(\frac{Y}{n} - p\right)^2 - \frac{z^2_{\alpha/2} \cdot p(1-p)}{n}}_{} \leq 0$$

this is quadratic in p. Call it $H(p)$.



this interval will be our C.I.

- Let $\hat{p} = y/n$, then solve for the zeros of H:

$$\frac{\left(\hat{p} + \frac{z_{\alpha/2}^2}{2n}\right) \mp z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z_{\alpha/2}^2}{4n^2}}}{1 + \frac{z_{\alpha/2}^2}{n}}$$

- For large n, $\frac{z_{\alpha/2}^2}{2n}$, $\frac{z_{\alpha/2}^2}{4n^2}$, $\frac{z_{\alpha/2}^2}{n}$ will be small, making the two intervals approximately equal.

Ex: In the midterm question $y/n = 28/50 = 0.56$

① $0.56 \pm \underbrace{1.645}_{z_{0.05}} \times \sqrt{\frac{0.56 \times (1-0.56)}{50}} = (0.444521,\ 0.675479)$

② ... $= (0.444402,\ 0.669437)$

Ex: { upper bound on the proportion of defectives in item manufactory
lower bound on the proportion of voters who favor a particular candidate

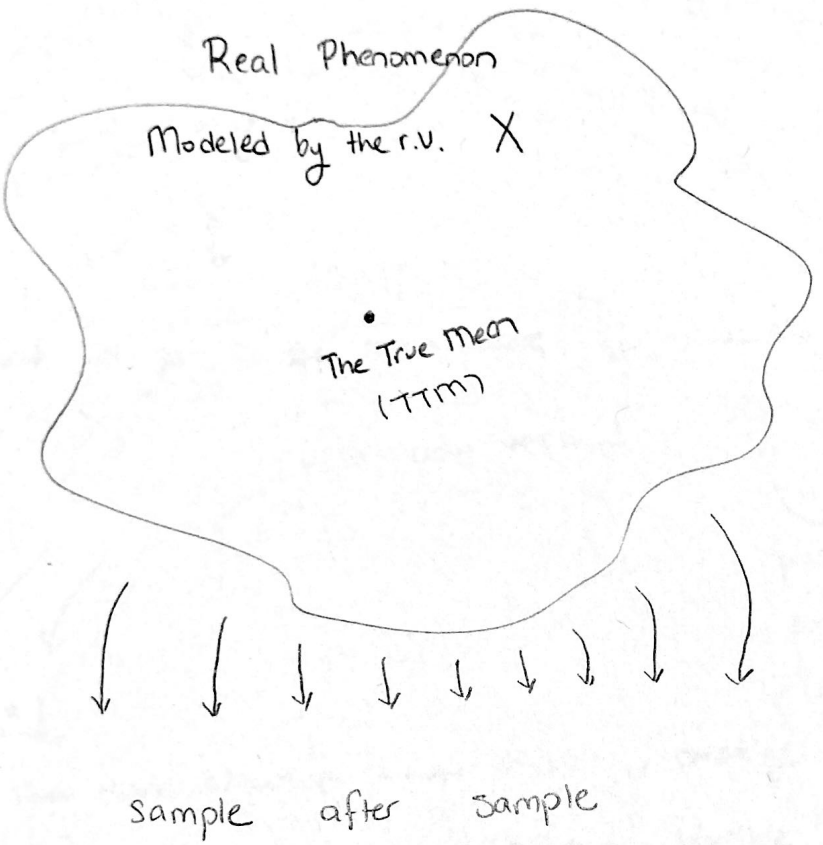
$$\left[0, \quad \frac{y}{n} + z_\alpha \cdot \sqrt{\frac{(y/n)\cdot(1-y/n)}{n}}\right]$$ (upper bound)

$$\left[\frac{y}{n} - z_\alpha \cdot \sqrt{\frac{(y/n)\cdot(1-y/n)}{n}}, \quad 1\right]$$ (lower bound)

HW Read on p 327 on how to compare two different "supposed" p's for the same interval probabilities.

HW 1 - 7

7.4 Sample Size



- How large should the sample size be to estimate TTM?

- If $Var(X) = 0$, then $n = 1$.

- Putting this silly example aside, the smaller $Var(X)$ the smaller the necessary n should be.

Ex: New method of teaching calculus

⇓

legitimacy?

⇓

based on μ: the mean score for students who learned calculus with new method

past experience suggests $\sigma = 15$

$n = 1$

- Isaac Newton → 100/100, amazing method!
- Uzumaki Naruto → 40/100 horrible method!!

So the sample size must be increased. To what?

100?

1000?

1,000,000?

- This is now a decision problem. As all such problems it must adhere to certain, well-defined restrictions.

- We know we won't get μ exactly (unless the underlying model is trivial). We could ask for μ to be around a certain neighborhood of the observed sample mean, with certain <u>confidence</u> level:

  (i) want μ to be in $\bar{X} \mp 1$, with 95% confidence

  (ii) want μ to be in $\bar{X} \mp 2$, with 80% confidence

- $\bar{X} \sim N(\mu, \sigma^2/n) \rightsquigarrow$ 95% C.I.:

$$\bar{X} \mp z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

$z_{0.025} = 1.96, \quad \sigma = 15 \Rightarrow \cong 1 \Rightarrow \sqrt{n} \cong 1.96 \times 15 = 29.4$

$\Rightarrow n \cong 864.34$

- $n = 865$ students would be enough to achieve (i)

- For the condition (ii) to be satisfied, we'd need

$$z_{0.1} = 1.282, \quad \sigma = 15$$

$$1.282 \times \frac{15}{\sqrt{n}} \cong 2 \quad \Rightarrow \quad \sqrt{n} = 9.615$$

$$\Rightarrow \quad n \cong 92.45$$

- $n = 93$ students, more feasible to conduct, depending on the total student population.

- Now we know these benchmark values, assume we have more students, say $n = 145$, we can reduce the interval size or increase the confidence level, or a combination of the two.

- Further, as n is large enough, we may also want to use s^2 instead of the "prior" σ^2; for example:

80% confidence

$s = 13.2$ $\rightsquigarrow$ $\bar{x} \pm 1.41$ $\searrow$ a smaller interval.

$n = 145$

The General Picture: $100 \cdot (1-\alpha)\%$ C.I for μ: $\bar{X} \mp z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$

- If we do not want the interval to be no longer than $\bar{X} \mp \epsilon$, then the smallest n that achieves this:

$$n = \left\lceil \frac{z^2_{\alpha/2} \cdot \sigma^2}{\epsilon^2} \right\rceil,$$

assuming σ^2 is known. This ϵ is called the maximum error of the estimate.

HW Read p331 for a further modification to choose an appropriate sample size when using s^2 instead of σ^2 and Example 7.4-2.

- Now, we'll look to apply similar ideas to proportions instead of means.

- Examples of proportions:
  - % labor force that is unemployed
  - % voters favoring a certain candidate.

- Estimating these unknown proportions might lead to important policy-making decisions by governments or private entities.

Short confidence intevals, high confidence levels

$\Rightarrow$

large sample size

Ex: Suppose that in the last year the unemployment rate has been about 8%. Before an important policy decision, we'd like to update our estimate.

criterion: 99% confident that the new estimate will be within 0.001 of the true p.

- Recall from section 7.3: $99\%\ \text{C.I.} \approx \frac{y}{n} \mp 2.576 \cdot \sqrt{\frac{(y/n)(1-y/n)}{n}}$

- We cannot know $\hat{p}/n$ before sampling, but it can be estimated by the prior amount, 0.08:

$$2.576 \cdot \sqrt{\frac{0.08 \times 0.92}{n}} \approx 0.001$$

$$\Rightarrow \quad n \approx \left(\frac{2.576}{0.001}\right)^2 \times 0.08 \cdot 0.92$$

$$\Rightarrow \quad n \approx 488,393.11$$

$$\Rightarrow \quad n \approx 488,394$$

- Error $0.001 \longrightarrow 0.01$, $99\% \longrightarrow 98\%$ confidence

$$\sqrt{n} \cong \frac{2.326}{0.01} \times \sqrt{0.0736} \quad \Rightarrow \quad n \cong 3,981.96$$

much more reasonable

Ex: "51% of the voters seem to favor candidate Z."

- $n = 100 \longrightarrow [0.41, 0.61]$ ×
- $n = 2000 \longrightarrow [0.49, 0.53]$ ×
- $n = 10,000 \longrightarrow [0.50, 0.52]$ ✓

- Similar to what we did for means, we define

$$\epsilon = z_{\alpha/2} \cdot \sqrt{\frac{\hat{p} \cdot (1-\hat{p})}{n}}$$

to be the maximum error of the point estimate $\hat{p} = y/n$.

- In the previous example we estimated this estimator by 0.08.

What if we do not have such prior knowledge?

- Consider the function $g(p) = p(1-p) = p - p^2$

$g'(p) = 1 - 2p = 0 \Rightarrow p^* = 1/2$ is a critical pt.

$g''(p^*) = -2 < 0 \Rightarrow p^* = 1/2$ a local max.

- Further, as g is a quadratic polynomial that's concave down it has a unique (and thus global) maximum, this $p^* = 1/2$.

- Long story short, we can always bound $p \cdot (1-p) \leq 1/4$.

- Then, $$n = \frac{z_{\alpha/2}^2 \cdot \hat{p}\cdot(1-\hat{p})}{\epsilon^2} \leq \frac{z_{\alpha/2}^2}{4\epsilon^2}$$

Ex: Candidate wants to asses initial support among the voters.

$$p \gtrapprox 0.15 \Rightarrow \text{enter}$$

$$p \ll 15 \Rightarrow \text{do not enter}$$

Criterion: 95% confident that p will be in $y/n \mp 0.03$.

- No prior knowledge on p: $$n \approx \frac{z_{0.025}^2}{4\times 0.03^2} \cong 1067.11$$

$$n \cong 1068$$

- They sampled 1068 voters, $y = 214$ supported the candidate.

$$\hat{p} = \frac{214}{1068} \cong 0.20$$

95% C.I.: $$0.20 \mp 1.96 \times \sqrt{\frac{0.2\times 0.8}{1068}} \rightsquigarrow 0.20 \mp 0.024$$

smaller than $\epsilon = 0.03$?

- This is because we used $p^* = 0.5$ instead of the more accurate 0.2.

- What if the population size vs desired sample size are comparable?

- The sample size could be adjusted significantly.

HW Work on p 334 - 336 & Example 7.4-5.

$N = 3000$

$\downarrow$

$n = 1068$ (max p based)

$\downarrow$ adj.

$n = 788$

HW 1 - 10

7.5 Distribution-Free Confidence Intervals for Percentiles

- Median, quartiles & percentiles in general provide a good means to understand the underlying distribution.

- We saw in Chapter 6 how to estimate these using sample percentiles, which were calculated using order statistics. These were point estimates.

- In this section we'll see how to construct confidence intervals for the estimation of percentiles.

- Our approach will be direct calculation based one, without any distributional assumption, which makes it nonparametric.

- Such confidence intervals are called distribution-free confidence intervals.

Ex: Consider the order statistics of a random sample of size $n=5$:

$$Y_1 < Y_2 < Y_3 < Y_4 < Y_5$$

For the median, $m = \pi_{0.5}$, $(n+1).p = (5+1).0.5 = 3$, so,

$$\tilde{\pi}_{0.5} = Y_3$$ was the point estimator for m.

- As $n=5$ small, an interval estimator would be much better suited, as there's a high chance that the point estimator will be significantly off.

- We stated we won't assume any distributional structure (Normal, t, χ^2, etc.) other than the distribution being a cts. one. This means we cannot construct the C.I. the way we did previously.

- Instead, we'll turn to what we can estimate: Y_i's.

- Let's determine $P(Y_1 < m < Y_5)$

this is a difficult problem if you attempt to solve it via the joint pdf of Y_1, Y_5:

- we don't know the underlying pdf/CDF, may estimate via empirical
- calculations would be very involved.

- Clever way out: consider the Bernoulli trials: $\{X_i < m\}$

by definition of m, this is like a fair coin toss: $P(X_i < m) = 0.5$

- The event $\{Y_1 < m < Y_5\}$ iff at least one success
but
not five successes

$$P(Y_1 < m < Y_5) = \sum_{k=1}^{4} \binom{5}{k} \cdot (1/2)^k (1-1/2)^{5-k}$$

$$= \underbrace{\sum_{k=0}^{5} \binom{5}{k} (1/2)^k \cdot (1-1/2)^{5-k}}_{=1} - \underbrace{\left(\frac{1}{2}\right)^5}_{k=0} - \underbrace{\left(\frac{1}{2}\right)^5}_{k=5}$$

$$= 1 - \frac{1}{2^5} - \frac{1}{2^5} = \frac{30}{32} = \frac{15}{16} = 1 - \frac{1}{2^4}$$

$$\approx 0.94$$

- Once $Y_1 = y_1$ and $Y_5 = y_5$ are realized, (y_1, y_5) will be a 94% C.I. for the true median.

- We will do two generalizations:
  - to an arbitrary sample size
  - to an arbitrary percentile, π_p, $p \in [0,1]$.

- First, assume $X_1, \ldots, X_n$ is our random sample and $Y_1 < \ldots < Y_n$, the corresponding order statistics. Then,

$$P(Y_1 < m < Y_n) = \sum_{k=1}^{n-1} \binom{n}{k} \cdot \left(\tfrac{1}{2}\right)^k \left(\tfrac{1}{2}\right)^{n-k} = 1 - \left(\tfrac{1}{2}\right)^n - \left(\tfrac{1}{2}\right)^n$$

$$= 1 - \left(\tfrac{1}{2}\right)^{n-1}$$

- As $n \uparrow$, the confidence level gets close to 100% exponentially fast.

- Sounds good? The interval (y_1, y_n) also quickly gets large, to the degree that it's not practical.

- 99.9999% C.I. for the mean age in our class: $[-1, 1{,}000]$

$\{$

may be true, but not informative!

- Instead of (Y_1, Y_n), we might want to use (Y_2, Y_{n-1}) or (Y_3, Y_{n-2}), or even "smaller" intervals in order to trade confidence levels for more reasonably shorter intervals.

- In general for $1 \leq i < j \leq n$,

$$P(Y_i < m < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} \cdot (1/2)^k \cdot (1/2)^{n-k}$$

$$= 1-\alpha \; ???$$

- We'll set a fixed $\alpha \in [0,1]$, then $i \uparrow$ $j \downarrow$ as necessary to reach to the desired level of confidence, $1-\alpha$.

- At this point we can utilize two approaches:

  (i) direct calculation (gets difficult)

  (ii) Normal approximation to Binomial probabilities, if n is large enough.

Ex: $n=9$ fish captured off the New England coast. Before the sample is drawn,

$$P(Y_2 < m < Y_8) = \sum_{k=2}^{7} \binom{9}{k} \cdot (1/2)^k (1/2)^{9-k}$$

$$= (1/2)^9 \cdot \underbrace{\sum_{k=2}^{7} \binom{9}{k}}$$

$$= 2^9 - (1+9+9+1) = 2^9 - 20$$

$$\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{n} = 2^n$$

$$= 1 - \frac{20}{2^9} = 1 - \frac{20}{512} \cong 0.9609375$$

- Or, you can make use of the Binomial CDF provided in Table II on p494.,

- Say we have the following realized order statistics:

$$15.5 < \underset{\uparrow}{19.0} < 21.2 < 21.7 < 22.8 < 27.6 < 29.3 < \underset{\uparrow}{30.1} < 32.5$$

y_2 $\qquad$ y_8

- So the interval $(19, 30.1)$ is a 96.1% C.I. for the true median m.

table 7.5-1

- Some specific examples of probabilities of the form $P(Y_i < m < Y_j)$ are provided in p 339 for the ease of calculation.

↓

Table 7.5-1

↓

Choices of $(i, n-i+1)$, $P > 90\%$ and as close to 95% as possible

$n=9$	$(2, 8)$	$P(Y_2 < m < Y_8) = 0.9610$
$n=14$	$(4, 11)$	$P(Y_4 < m < Y_{11}) = 0.9426$
$n=16$	$(5, 12)$	$P(Y_5 < m < Y_{12}) = 0.9232$
$n=20$	$(6, 15)$	$P(Y_6 < m < Y_{15}) = 0.9586$

- Now let's see how we can utilize Normal approx.

Ex: Let $n=16$, instead of calculating $P(Y_5 < m < Y_{12}) = 0.9232$,

let's Normally approximate it:

We want $1-\alpha = P(Y_5 < m < Y_{12}) = \sum_{k=5}^{11} \binom{16}{k} \cdot (1/2)^k (1/2)^{16-k}$

This is, for $W \sim \text{Binomial}(16, 1/2)$, $= P(5 \leq W \leq 11)$

$E[W] = 16 \times \frac{1}{2} = 8$ $= P(4.5 < W < 11.5)$ (c.c.)

$\text{Var}(W) = 16 \times \frac{1}{2} \times \frac{1}{2} = 4$

$$1-\alpha = P(Y_5 < m < Y_{12}) = P(4.5 < W < 11.5)$$

$$= P\left(\frac{4.5-8}{2} < \frac{W-8}{2} < \frac{11.5-8}{2}\right)$$

$\frac{\bar{X}-\mu}{\sigma}$

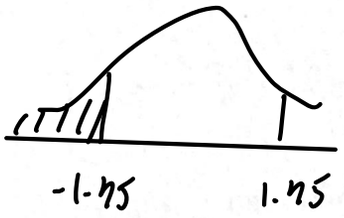
$$\phi(\quad) \underset{(CLT)}{\cong} P(-1.75 < Z < 1.75), \text{ for } Z \sim N(0,1)$$

$$= \phi(1.75) - \phi(-1.75) \longrightarrow 1 - \phi(1.75)$$

$$= 2\phi(1.75) - 1$$

$$= 0.9599 - 0.0401$$

$$= 0.9198$$



↓ compare to the true probability:

0.9232

- For $n=16$, this is good enough. This approach will shine when n is very large, exact calculations difficult & CLT approximation is very good.

- Now we will estimate π_p, p any real from 0 to 1, the same way we did $m = \pi_{0.5}$. Noting $P(X < \pi_p) = p$, for $1 \le i < j \le n$ we want

$$1-\alpha = P(Y_i < \pi_p < Y_j) = \sum_{k=i}^{j-1} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k},$$

- as we want at least i successes for $Y_i < \pi_p$, but fewer than j successes.

HW Go over Examples 7.5-2, 7.5-3

importance of stem-leaf display to manually find the order statistic values.

HW Read the last paragraph comparing the methods of 7.1 - 7.3 (parametric) vs 7.5. Intervals for mean vs median should be comparable, but the nonparametric intervals tend to be larger as they assume less, making them more robust.

HW 1 - 9