

Order statistics, quantiles & resampling

(Module 9)

Statistics (MAST20005) & Elements of Statistics (MAST90058)

Semester 2, 2020

Contents

1	Order statistics	1
1.1	Introduction	1
1.2	Sampling distribution	3
2	Quantiles	6
2.1	Definitions	6
2.2	Asymptotic distribution	10
2.3	Confidence intervals for quantiles	10
3	Resampling methods	12

Aims of this module

- Go back to **order statistics** and **sample quantiles**
- More detailed definitions
- Derive **sampling distributions** and construct **confidence intervals**
- See examples of CIs that are **not** of the form $\hat{\theta} \pm \text{se}(\hat{\theta})$
- Learn some more distribution-free methods
- See how to use computation to avoid mathematical derivations

Unifying theme

- Use the data ‘directly’ rather than via assumed distributions
- Use the **sample cdf** and related summaries (such as order statistics)

1 Order statistics

1.1 Introduction

Definition (recap)

- Sample: X_1, \dots, X_n
- Arrange them in increasing order:

$$\begin{aligned} X_{(1)} &= \text{Smallest of the } X_i \\ X_{(2)} &= \text{2nd smallest of the } X_i \\ &\vdots \\ X_{(n)} &= \text{Largest of the } X_i \end{aligned}$$

- These are called the *order statistics*

$$X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(n)}$$

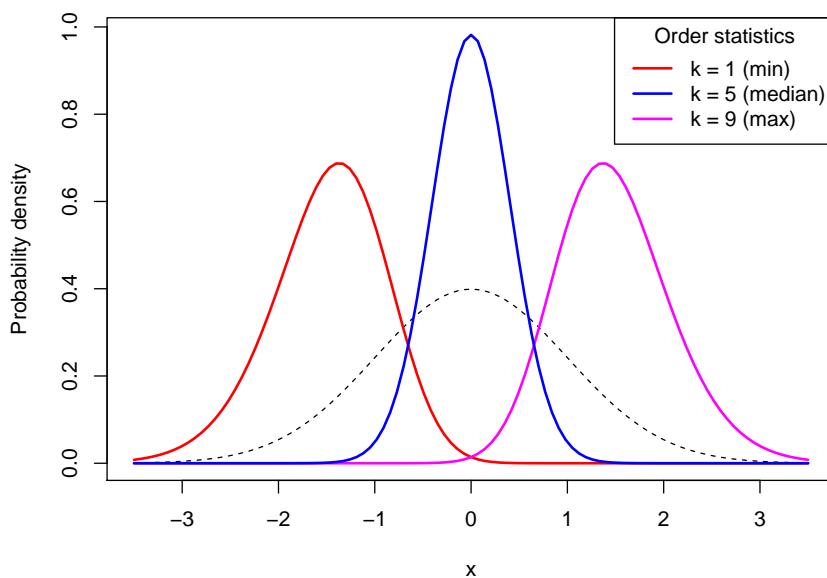
- $X_{(k)}$ is called the *kth order statistic* of the sample
- $X_{(1)}$ is the *minimum* or *sample minimum*
- $X_{(n)}$ is the *maximum* or *sample maximum*

Motivating example

- Take iid samples $X \sim N(0, 1)$ of size $n = 9$
- What can we say about the order statistics, $X_{(k)}$?
- Simulated values:

	[,1]	[,2]	[,3]	[,4]	[,5]	
[1,]	-0.76	-1.94	-1.32	-0.85	-1.96	<-- Minimum
[2,]	-0.32	-0.17	-0.53	-0.30	-0.98	
[3,]	-0.23	0.06	-0.44	0.14	-0.83	
[4,]	0.05	0.18	-0.10	0.25	-0.63	
[5,]	0.08	0.76	0.17	0.35	-0.47	<-- Median
[6,]	0.18	0.96	0.26	0.68	0.05	
[7,]	0.27	1.07	0.60	0.69	0.34	
[8,]	0.73	1.42	0.66	1.13	1.26	
[9,]	0.91	1.77	1.93	1.98	1.26	<-- Maximum

Standard normal distribution, n = 9



1.2 Sampling distribution

Example (triangular distribution)

- Random sample: X_1, \dots, X_5 with pdf $f(x) = 2x$, $0 < x < 1$
- Calculate $\Pr(X_{(4)} \leq 0.5)$
- Occurs if at least four of the X_i are less than 0.5,

$$\begin{aligned}\Pr(X_{(4)} \leq 0.5) &= \Pr(\text{at least 4 } X_i\text{'s less than 0.5}) \\ &= \Pr(\text{exactly 4 } X_i\text{'s less than 0.5}) \\ &\quad + \Pr(\text{exactly 5 } X_i\text{'s less than 0.5})\end{aligned}$$

- This is a binomial with 5 trials and probability of success given by

$$\Pr(X_i \leq 0.5) = \int_0^{0.5} 2x \, dx = [x^2]_0^{0.5} = 0.5^2 = 0.25$$

- So we have,

$$\Pr(X_{(4)} \leq 0.5) = \binom{5}{4} 0.25^4 0.75 + 0.25^5 = 0.0156$$

- More generally we have,

$$F(x) = \Pr(X_i \leq x) = \int_0^x 2t \, dt = [t^2]_0^x = x^2$$

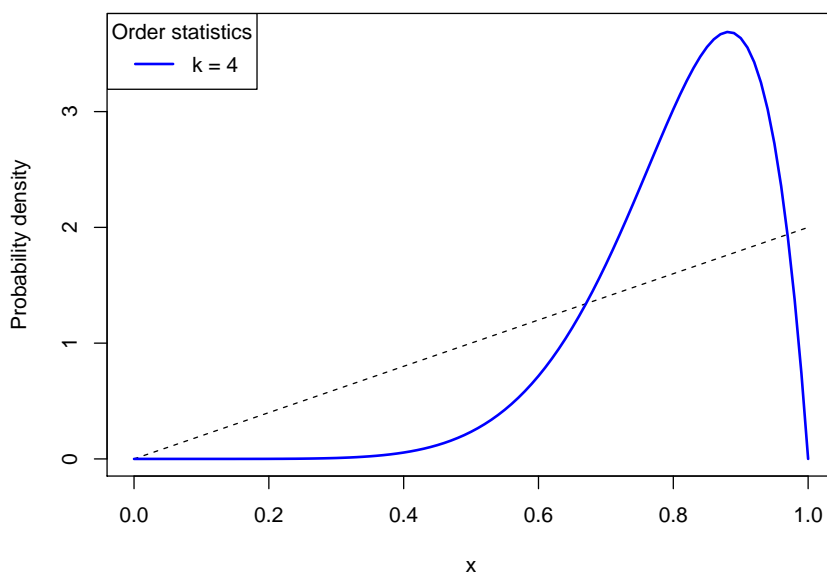
$$G(x) = \Pr(X_{(4)} \leq x) = \binom{5}{4} (x^2)^4 (1 - x^2) + (x^2)^5$$

- Taking derivatives gives the pdf,

$$\begin{aligned}g(x) = G'(x) &= \binom{5}{4} 4(x^2)^3 (1 - x^2)(2x) \\ &= 4 \binom{5}{4} F(x)^3 (1 - F(x)) f(x)\end{aligned}$$

since we know that $F(x) = x^2$.

Triangular distribution, n = 5



Distribution of $X_{(k)}$

- Sample from a continuous distribution with cdf $F(x)$ and pdf $f(x) = F'(x)$.
- The cdf of $X_{(k)}$ is,

$$\begin{aligned} G_k(x) &= \Pr(X_{(k)} \leq x) \\ &= \sum_{i=k}^n \binom{n}{i} F(x)^i (1 - F(x))^{n-i} \end{aligned}$$

- Thus the pdf of $X_{(k)}$ is,

$$\begin{aligned} g_k(x) &= G'_k(x) = \sum_{i=k}^n i \binom{n}{i} F(x)^{i-1} (1 - F(x))^{n-i} f(x) \\ &\quad + \sum_{i=k}^{n-1} (n-i) \binom{n}{i} F(x)^i (1 - F(x))^{n-i-1} (-f(x)) \\ &= k \binom{n}{k} F(x)^{k-1} (1 - F(x))^{n-k} f(x) \\ &\quad + \sum_{i=k+1}^n i \binom{n}{i} F(x)^{i-1} (1 - F(x))^{n-i} f(x) \\ &\quad - \sum_{i=k}^{n-1} (n-i) \binom{n}{i} F(x)^i (1 - F(x))^{n-i-1} f(x) \end{aligned}$$

- But

$$i \binom{n}{i} = \frac{n!}{(i-1)!(n-i)!} = n \binom{n-1}{i-1}$$

and similarly

$$(n-i) \binom{n}{i} = \frac{n!}{i!(n-i-1)!} = n \binom{n-1}{i}$$

which allows some cancelling of terms.

- For example, the first term of the first summation is,

$$\begin{aligned} (k+1) \binom{n}{k+1} F(x)^k (1 - F(x))^{n-k-1} f(x) \\ = n \binom{n-1}{k} F(x)^k (1 - F(x))^{n-k-1} f(x) \end{aligned}$$

- The first term of the second summation is,

$$\begin{aligned} (n-k) \binom{n}{k} F(x)^k (1 - F(x))^{n-k-1} f(x) \\ = n \binom{n-1}{k} F(x)^k (1 - F(x))^{n-k-1} f(x) \end{aligned}$$

- These cancel, and similarly the other terms do as well.
- Hence, the pdf simplifies to,

$$g_k(x) = k \binom{n}{k} F(x)^{k-1} (1 - F(x))^{n-k} f(x)$$

- Special cases: minimum and maximum,

$$\begin{aligned} g_1(x) &= n (1 - F(x))^{n-1} f(x) \\ g_n(x) &= n F(x)^{n-1} f(x) \end{aligned}$$

- Also:

$$\begin{aligned} \Pr(X_{(1)} > x) &= (1 - F(x))^n \\ \Pr(X_{(n)} \leq x) &= F(x)^n \end{aligned}$$

Alternative derivation of the pdf of $X_{(k)}$

- Heuristically,

$$\Pr(X_{(k)} \approx x) = \Pr(x - \frac{1}{2}dy < X_{(k)} \leq x + \frac{1}{2}dy) \approx g_k(x) dy$$

- Need to observe X_i such that:

- $k - 1$ are in $(-\infty, x - \frac{1}{2}dy]$
- One is in $(x - \frac{1}{2}dy, x + \frac{1}{2}dy]$
- $n - k$ are in $(x + \frac{1}{2}dy, \infty)$

- Trinomial distribution (3 outcomes), event probabilities:

$$\begin{aligned}\Pr(X_i \leq x - \frac{1}{2}dy) &\approx F(x) \\ \Pr(x - \frac{1}{2}dy < X_i \leq x + \frac{1}{2}dy) &\approx f(x) dy \\ \Pr(X_i > x + \frac{1}{2}dy) &\approx 1 - F(x)\end{aligned}$$

- Putting these together,

$$g_k(x) dy \approx \frac{n!}{(k-1)! 1! (n-k)!} F(x)^{k-1} (1 - F(x))^{n-k} f(x) dy$$

- Dividing both sides by dy gives the pdf of $X_{(k)}$

Example (boundary estimate)

- $X_1, \dots, X_4 \sim \text{Unif}(0, \theta)$
- Likelihood is

$$L(\theta) = \begin{cases} \left(\frac{1}{\theta}\right)^4 & 0 \leq x_i \leq \theta, \quad i = 1, \dots, 4 \\ 0 & \text{otherwise (i.e. if } \theta < x_i \text{ for some } i) \end{cases}$$

- Maximised when θ is as small as possible, so $\hat{\theta} = \max(X_i) = X_{(4)}$
- Now,

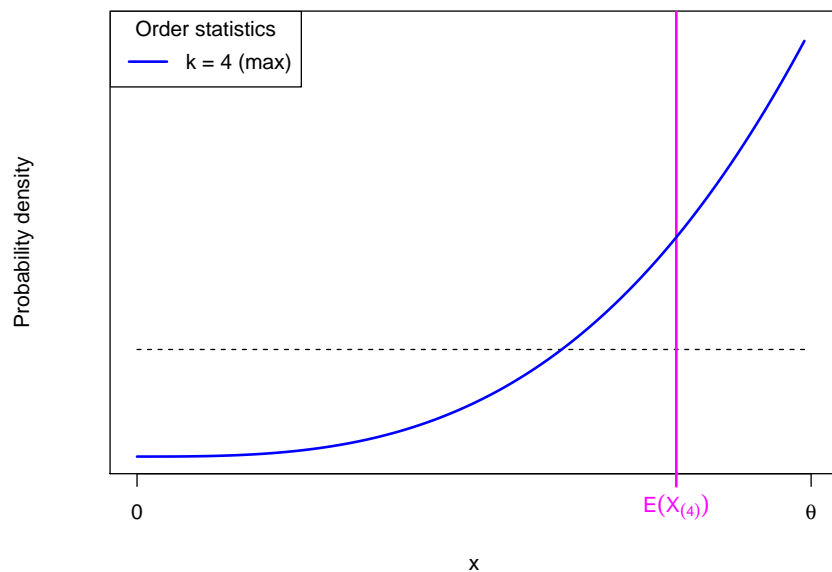
$$g_4(x) = 4 \left(\frac{x}{\theta}\right)^3 \left(\frac{1}{\theta}\right) = \frac{4x^3}{\theta^4}, \quad 0 \leq x \leq \theta$$

- Then,

$$\mathbb{E}(X_{(4)}) = \int_0^\theta x \frac{4x^3}{\theta^4} dx = \left[\frac{4x^5}{5\theta^4} \right]_0^\theta = \frac{4}{5}\theta$$

- So the MLE $X_{(4)}$ is biased
- (But $\frac{5}{4}X_{(4)}$ is unbiased)

Uniform distribution, n = 4



- Deriving a one-sided CI for θ based on $X_{(4)}$:

- For a given $0 < c < 1$, show that,

$$1 - c^4 = \Pr(c\theta < X_{(4)} < \theta) = \Pr(X_{(4)} < \theta < X_{(4)}/c)$$

- Thus, a $100 \cdot (1 - c^4)\%$ confidence interval for θ is $(x_{(4)}, x_{(4)}/c)$
- Letting $c = \sqrt[4]{0.05} = 0.47$, we have a 95% confidence interval from $x_{(4)}$ to $2.11x_{(4)}$

2 Quantiles

2.1 Definitions

Population quantiles

- Informally, a **quantile** is a number that divides the range of a random variable based on the probabilities on either side.
- The p -quantile, π_p , of a continuous probability distribution with cdf F has the property:

$$p = F(\pi_p) = \Pr(X \leq \pi_p)$$

So, we can define it by the inverse cdf:

$$\pi_p = F^{-1}(p)$$

- More general definition (also works for discrete variables): the p -quantile is the smallest value π_p such that $p \leq F(\pi_p)$
- The most commonly used quantile is the *median*, $\pi_{0.5}$, often referred to simply as m
- Also the first and third *quartiles*, $\pi_{0.25}$ and $\pi_{0.75}$

Sample quantiles

- Want a statistic which estimates π_p
- There are many ways to do this
- R implements 9 different definitions!

- See `help(quantile)`
- Previously mentioned two of these...

‘Type 6’ quantiles

- Definition:

$$\hat{\pi}_p = x_{(k)}, \quad \text{where } p = \frac{k}{n+1}$$

- Linear interpolation otherwise
- Motivated by the following relationship (see later):

$$\mathbb{E}(F(X_{(k)})) = \frac{k}{n+1}$$

- We used this previously for QQ plots

‘Type 7’ quantiles

- Definition:

$$\hat{\pi}_p = x_{(k)}, \quad \text{where } p = \frac{k-1}{n-1}$$

- Linear interpolation otherwise
- Motivated by the following relationship (see later):

$$\text{mode}(F(X_{(k)})) = \frac{k-1}{n-1}$$

- This is the default in R (`quantile` function)

‘Type 1’ quantiles

- Can also apply the general quantile definition to the sample cdf:

$$\hat{\pi}_p = x_{(\lceil np \rceil)}$$

- The ceiling function, $\lceil b \rceil$, is the smallest integer not less than b
- In other words,

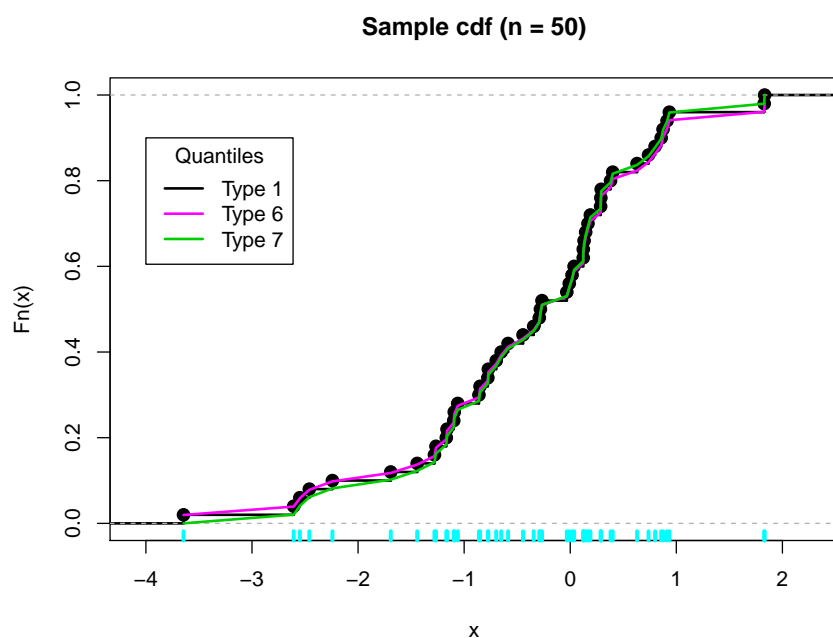
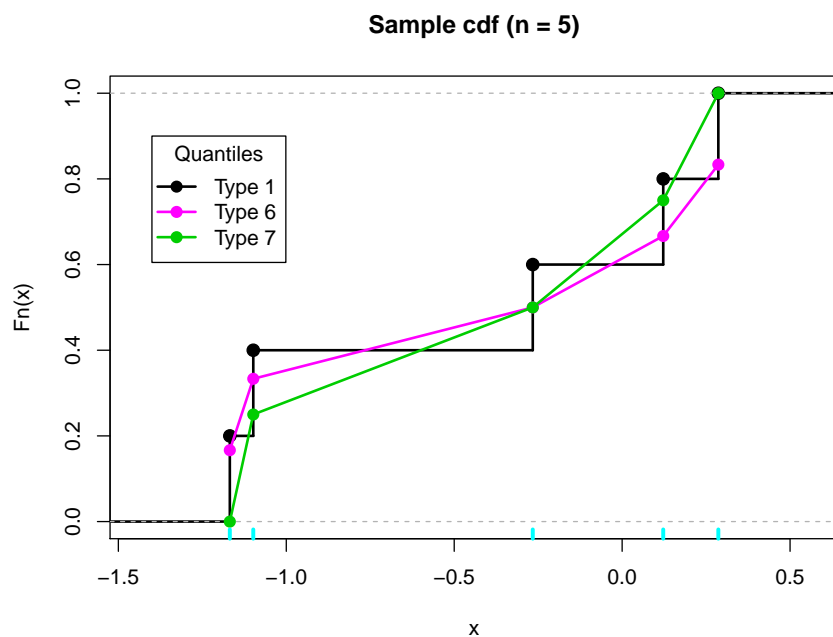
$$\hat{\pi}_p = x_{(k)}, \quad \text{if } \frac{k-1}{n} < p \leq \frac{k}{n}$$

- Reminder: the *sample cdf* is

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^n I(x_i \leq x)$$

Differences in definitions

- Different definitions imply different estimators for the cdf
- For large sample sizes, differences are negligible



Distribution on the cdf scale

- Reminder: for a continuous distribution, $F(X) \sim \text{Unif}(0, 1)$
- Proof: for $0 \leq w \leq 1$,

$$G(w) = \Pr(F(X) \leq w) = \Pr(X \leq F^{-1}(w)) = F(F^{-1}(w)) = w$$

so the density is

$$g(w) = G'(w) = 1, \quad 0 \leq w \leq 1$$

so $F(X) \sim \text{Unif}(0, 1)$.

- Since F is non-decreasing, we have

$$F(X_{(1)}) < F(X_{(2)}) < \dots < F(X_{(n)})$$

- So $W_i = F(X_{(i)})$ are order statistics from a $\text{Unif}(0, 1)$ distribution

- The cdf is $G(w) = w$, for $0 < w < 1$
- So the pdf of k th order statistic $W_k = F(X_{(k)})$ is

$$g_k(w) = k \binom{n}{k} w^{k-1} (1-w)^{n-k}$$

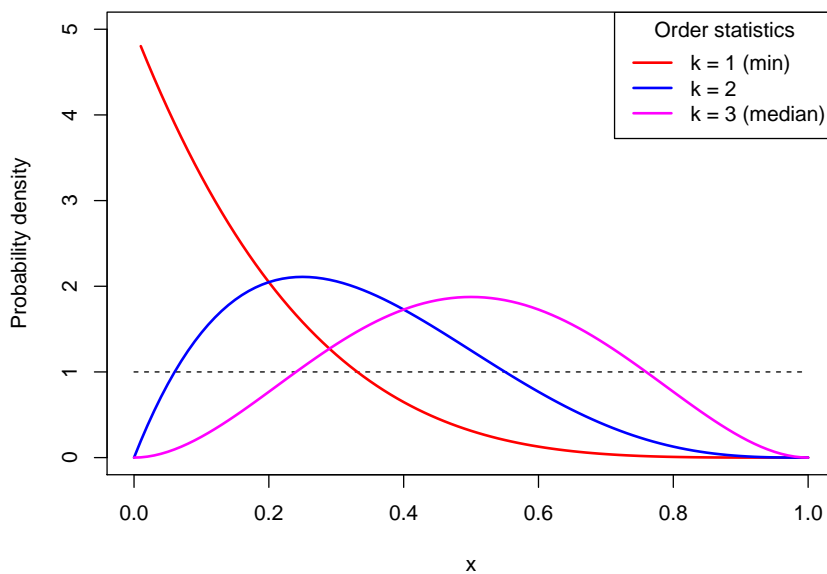
- This is a beta distribution,

$$F(X_k) \sim \text{Beta}(k, n - k + 1)$$

- We can derive that:

$$\begin{aligned} \mathbb{E}(W_k) &= \frac{k}{n+1} \\ \text{mode}(W_k) &= \frac{k-1}{n-1} \end{aligned}$$

Uniform distribution, n = 5



Defining the estimators

- How does this relate to the definitions of the estimators?
- Consider:

$$\begin{aligned} \Pr(X \leq X_{(k)}) &= F(X_{(k)}) \\ \Pr(X \leq \pi_p) &= F(\pi_p) = p \end{aligned}$$

- Have $F(X_{(k)})$ probability to the left of $X_{(k)}$, need p probability to the left π_p
- Just need to relate them
- $F(X_{(k)})$ is the (random!) area to the left $X_{(k)}$
- We know its distribution, so can summarise it
- For example, $\mathbb{E}(F(X_{(k)})) = k/(n+1)$
- This suggests $X_{(k)}$ can be an estimator of π_p where $p = k/(n+1)$
- So, define $\hat{\pi}_p = X_{(k)}$ where $p = k/(n+1)$
- For other values of p , linearly interpolate

Sample median

- The *sample median* is

$$\hat{m} = \begin{cases} X_{((n+1)/2)} & \text{when } n \text{ is odd} \\ \frac{1}{2} (X_{(n/2)} + X_{((n/2)+1)}) & \text{when } n \text{ is even} \end{cases}$$

- Consistent with most definitions of the sample quantiles (not type 1!)

2.2 Asymptotic distribution

Asymptotic distribution

- For large sample sizes, it can be shown that

$$\hat{\pi}_p \approx N\left(\pi_p, \frac{p(1-p)}{nf(\pi_p)^2}\right)$$

where f is the pdf of the population distribution

- The median, $\hat{M} = \hat{\pi}_{0.5}$, is convenient special case,

$$\hat{M} \approx N\left(m, \frac{1}{4nf(m)^2}\right)$$

Example (normal distribution)

- Random sample: $X \sim N(\mu, \sigma^2)$ of size n
- Compare \bar{X} and \hat{M} as estimators of μ
- Already know,

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- Now we also know,

$$\hat{M} \approx N\left(m, \frac{1}{4nf(m)^2}\right)$$

- Note that $m = \mu$ and,

$$f(m) = f(\mu) = \frac{1}{\sigma\sqrt{2\pi}}$$

- This gives,

$$\hat{M} \approx N\left(\mu, \frac{\pi}{2} \frac{\sigma^2}{n}\right)$$

- Does the $\pi/2$ look familiar?
- ... problem 3, week 2!
- The sample mean, \bar{X} , is a more *efficient* estimator of μ than the sample median, \hat{M}
- In other scenarios, it can be the other way around

2.3 Confidence intervals for quantiles

Confidence intervals for quantiles

- Can we construct **distribution-free** CIs for quantiles?
- Can do so based on order statistics
- Procedure is the ‘inverse’ of the sign test

Example (CI for median)

- Take iid samples X_1, \dots, X_5
- $X_{(3)}$ is an estimator of the median $m = \pi_{0.5}$
- For the median to be between $X_{(1)}$ and $X_{(5)}$ must have at least one $X_i < m$ but not five $X_i < m$
- If the distribution is continuous, $\Pr(X < m) = 0.5$
- Let W be the number of $X_i < m$, then $W \sim \text{Bi}(5, 0.5)$ and

$$\begin{aligned}\Pr(X_{(1)} < m < X_{(5)}) &= \Pr(1 \leq W \leq 4) \\ &= \sum_{k=1}^4 \binom{5}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{5-k} \\ &= 1 - 0.5^5 - 0.5^5 = \frac{15}{16} \approx 0.94\end{aligned}$$

- So $(x_{(1)}, x_{(5)})$ is a 94% confidence interval for m

Confidence intervals for the median

- In general, want i and j so that, to the closest possible extent,

$$\begin{aligned}\Pr(X_{(i)} < m < X_{(j)}) &= \Pr(i \leq W \leq j-1) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{n-k} \approx 1 - \alpha\end{aligned}$$

- Need to use computed binomial probabilities (e.g. R) to determine i and j
- Or use the normal approximation to the binomial
- Note that these confidence intervals do not arise from pivots and cannot achieve 95% confidence exactly

Example (lengths of fish)

- Lengths of 9 fish (in cm), in ascending order:
15.5, 19.0, 21.2, 21.7, 22.8, 27.6, 29.3, 30.1, 32.5
- Now,

$$\Pr(X_{(2)} < m < X_{(8)}) = \sum_{k=2}^7 \binom{9}{k} \left(\frac{1}{2}\right)^k \left(\frac{1}{2}\right)^{9-k} = 0.9610$$

- In R:

```
> pbinom(7, size = 9, prob = 0.5) -  
+ pbinom(1, size = 9, prob = 0.5)  
[1] 0.9609375
```
- So a 96.1% confidence interval for m is (19.0, 30.1)

Confidence intervals for arbitrary quantiles

- Argument can be extended to any quantile and any order statistics,
- For example, the i th and j th,

$$\begin{aligned}1 - \alpha &= \Pr(X_{(i)} < \pi_p < X_{(j)}) \\ &= \Pr(i \leq W \leq j-1) \\ &= \sum_{k=i}^{j-1} \binom{n}{k} p^k (1-p)^{n-k}\end{aligned}$$

Example (income distribution)

- Incomes (in \$100's) for a sample of 27 people, in ascending order:
161, 169, 171, 174, 179, 180, 183, 184, 186,
187, 192, 193, 196, 200, 204, 205, 213, 221,
222, 229, 241, 243, 256, 264, 291, 317, 376
- Want to estimate the first quartile, $\pi_{0.25}$
- W is the number of the X 's below $\pi_{0.25}$
- $W \sim \text{Bi}(27, 0.25) \approx N(\mu = 27/4 = 6.75, \sigma^2 = 81/16)$
- This gives

$$\begin{aligned}\Pr(X_{(4)} < \pi_{0.25} < X_{(10)}) \\&= \Pr(4 \leq W \leq 9) \\&= \Pr(3.5 < W < 9.5) \quad (\text{continuity correction}) \\&= \Phi\left(\frac{9.5 - 6.75}{9/4}\right) - \Phi\left(\frac{3.5 - 6.75}{9/4}\right) \\&= 0.815\end{aligned}$$

- So (\$17 400, \$18 700) is an 81.5% CI for the first quartile

3 Resampling methods

Resampling

- What if maths is too hard?
- Try a *resampling* method
- Replaces mathematical derivation with brute force computation
- Used for approximating sampling distributions, standard errors, bias, etc.
- Sometimes work brilliantly, sometimes not at all

Bootstrap

- Most popular resampling method: the *bootstrap*
- Basic idea:
 - Use the sample cdf as an approximation to the true cdf
 - Simulate new data from the sample cdf
 - Equivalent to sampling with replacement from the actual data
- Use these *bootstrap samples* to infer sampling distributions of statistics of interest
- This is an advanced topic
- Only a 'taster' is presented...
- ...in the lab (week 10)