Recap

- **p-quantile of a population:** A value ξ_p such that

$$P(X < \xi_p) \le p \quad \text{and} \quad P(X > \xi_p) \le 1-p$$

- ξ_p is unique if X is continuous, but not when X is discrete. Uniqueness can be enforced by redefining ξ_p as

ξ_p = smallest x such that $F(x) \ge p$.

- Essentially there is p probability on the left of ξ_p.
- $p = 0.5$: $\xi_{0.5}$ = Median M
- $(\xi_{0.25}, \xi_{0.5}, \xi_{0.75}) = (Q_1, Q_2, Q_3)$ — quartiles.

- **p-quantile of a sample:** A number $\hat{\xi}_p$ that exceeds at most $100p\%$ of the sample (i.e., np observations) and is exceeded by at most $100(1-p)\%$ of the sample (i.e., $n(1-p)$ observations.)

- $(100 \times p)$-th percentile = p-th quantile
- Measures of location: Mean, Median, quantiles

↑ measures of "center"

$(\hat{Q}_1, \hat{Q}_2, \hat{Q}_3) = (\hat{\xi}_{0.25}, \hat{\xi}_{0.5}, \hat{\xi}_{0.75}) = ?$

Ex: (CPU data) Sample quartiles of the CPU data.

Recall: $\bar{X} = 48.2$

$n = 30$.

```
# > sort(cpu)
# [1]   9  15  19  22  24  25  30  34  35  35  36  36
37  38  42  43  46  48
# [19]  54  55  56  56  59  62  69  70  82  82  89 139
# >
```

$\hat{\xi}_{0.75}$ = 34 (15th) ; $\hat{\xi}_{0.75}$ = 59 (16th)

$\hat{\xi}_{0.5} = \frac{42+43}{2} = 42.5$

$\hat{\xi}_{0.25} = ?$ $\quad p = 0.25$ $\quad np = (30)(0.25) = 7.5$, $\quad n(1-p) = (30)(0.75) = 22.5$

From the dy: $\hat{\xi}_{0.25} = 34$

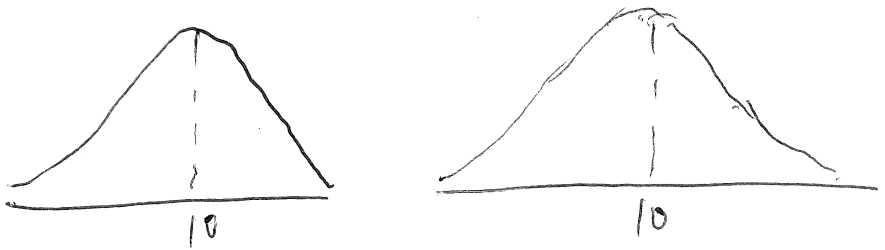
Similarly $\hat{\xi}_{0.75} = 59$.

"Measures of spread"

Population variance: $\sigma^2 = E((X-\mu)^2) = E(X^2) - \mu^2$; $\mu = E(X)$.

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$



Properties:

- s^2 estimates σ^2.
- $E[s^2] = \sigma^2$ for all σ^2 — unbiased
  → Verify in a HW.
- Measure of spread or variability
- Standard deviation (SD) = $\sqrt{\text{variance}}$
- Estimated standard error (SE) of $\bar{X}$ = Estimate of $\frac{\sigma}{\sqrt{n}} = \frac{s}{\sqrt{n}}$

Recall: $\text{var}[\bar{X}] = \frac{\sigma^2}{n}$

Ex: (CPU data)

$$s^2 = \frac{1}{30-1}\left\{(9-42.5)^2 + (15-42.5)^2 + \cdots + (139-42.5)^2\right\}$$

48.2 48.2 48.2

$= ?$

$s = \sqrt{s^2}$

NOTE:

- s^2 is greatly sensitive to outliers — not a good measure if outliers are present
- σ^2 is a good measure of spread if dist is

symmetric looking, but not when the dist. is skewed.

Interquartile range (IQR): — Alternative to σ^2 or s^2

Population: $IQR = Q_3 - Q_1$ → range of the middle 50% of the dist.

Sample: $\widehat{IQR} = \hat{Q}_3 - \hat{Q}_1$

Properties:

- Sample IQR estimates population IQR.
- Robust to outliers. — useful when outliers are present.
- '1.5 IQR' rule

Rule of thumb for "outlier" detection: An observation may be considered an "oulier" if it falls outside the interval from $\hat{Q}_1 - 1.5 * \widehat{IQR}$ to $\hat{Q}_3 + 1.5 * \widehat{IQR}$.

Ex: (CPU data): Estimated (or sample) IQR=? Could the observation 139 be an outlier?

$\widehat{IQR} = 59 - 34 = 25$.

check using this rule that 139 is indeed an outlier.

Which measure of spread to use — SD or IQR?:

dist. is skewed or outliers

population distn. or data distn is symmetric