University of Washington

STATISTICS

Elements of Statistical Methods

Population Attributes

(Ch 6)

Fritz Scholz

Spring Quarter 2010

February 14, 2010

Populations

We have introduced probability models, random variables and their distributions.

Random variables represent the numerical outcomes of a random experiment.

Randomness enters either through deliberate randomization on our part or through parsimonious modeling of the many uncontrollable factors that affect a numerical experimental outcome.

The distribution of a random variable is also referred to as the population.
The population notion arises from the fact that in many applications we sample a physical population (people, fish, manufactured parts, etc.) and observe a particular numerical characteristic for the sampled items.

The population consists of the measurements for all the items in the population.

Even if the population is typically finite, it often is so large that discrete distribution descriptions are not practical $\Longrightarrow$ continuous distributions/populations.

Characterizing Populations

In the face of the inherent experimental uncertainty, the best we can do is to give a complete population description.

It would allow us to provide various probabilistic statements about the population, given that deterministic statements are not feasible.

Unfortunately, knowing the full population is often not practical, since it involves getting the numerical characteristics for all population members. However, useful population approximations are still possible and will be discussed.

The more modest goal of characterizing basic aspects/properties of such populations is often more realistic.

Two main characteristics of any numerical population are the population center or middle and the population spread or dispersion.

The Population Center

For a normal population there is no question as to what should be the middle.

It is the parameter μ, the mean, the median and the mode.

We had interpreted the mean as the point of distribution balance.

The median ν is characterized by the property that $P(X \geq \nu) \geq 0.5 \leq P(X \leq \nu)$,

For the normal population we have $P(X \geq \mu) = P(X \leq \mu) = 0.5$ so that $\nu = \mu$.

The mode is that value at which the pdf or pmf has its highest value.

Again it is μ for $\mathcal{N}(\mu, \sigma^2)$. We will not dwell much on the mode.

One reason why there was no dispute concerning the distribution center

of the normal distribution is that its density is symmetric around μ, i.e.,

$$f(\mu - x) = f(\mu + x) \qquad \text{for all } x \in R$$

Symmetric Populations

For general continuous random variables with pdf f we say that its distribution or population is symmetric around some value θ whenever

$$f(\theta - x) = f(\theta + x) \qquad \text{for all } x \in R$$

Example: $X \sim \text{Uniform}(a, b)$ is symmetric around $(a+b)/2$.

More generally, a discrete or continuous r.v. X has a symmetric distribution if there exists a $\theta \in R$ such that $Y = X - \theta$ and $-Y = \theta - X$ have the same distribution. We write $Y \stackrel{\mathcal{D}}{=} -Y$ to express equality in distribution.

Thus the probabilities or pdf at $Y = y$ and $-Y = y$ or $Y = -y$ are the same for any $y \in R$ and should balance each other so that

$$EY = 0 \implies E(X - \theta) = EX - \theta = 0 \implies EX = \theta$$

provided that EY exists and is finite. There are pathological examples of symmetric distributions for which the expectation does not exist.

Quantiles

Definition: For a continuous r.v. X and for $\alpha \in (0,1)$ any number $q = q_\alpha = q(X;\alpha)$ with

$$P(X < q) = \alpha \qquad \text{and} \qquad P(X > q) = 1 - \alpha$$

is called an α-quantile of X or the associated distribution or population.

The α-quantile may not be unique, see later slide for example.

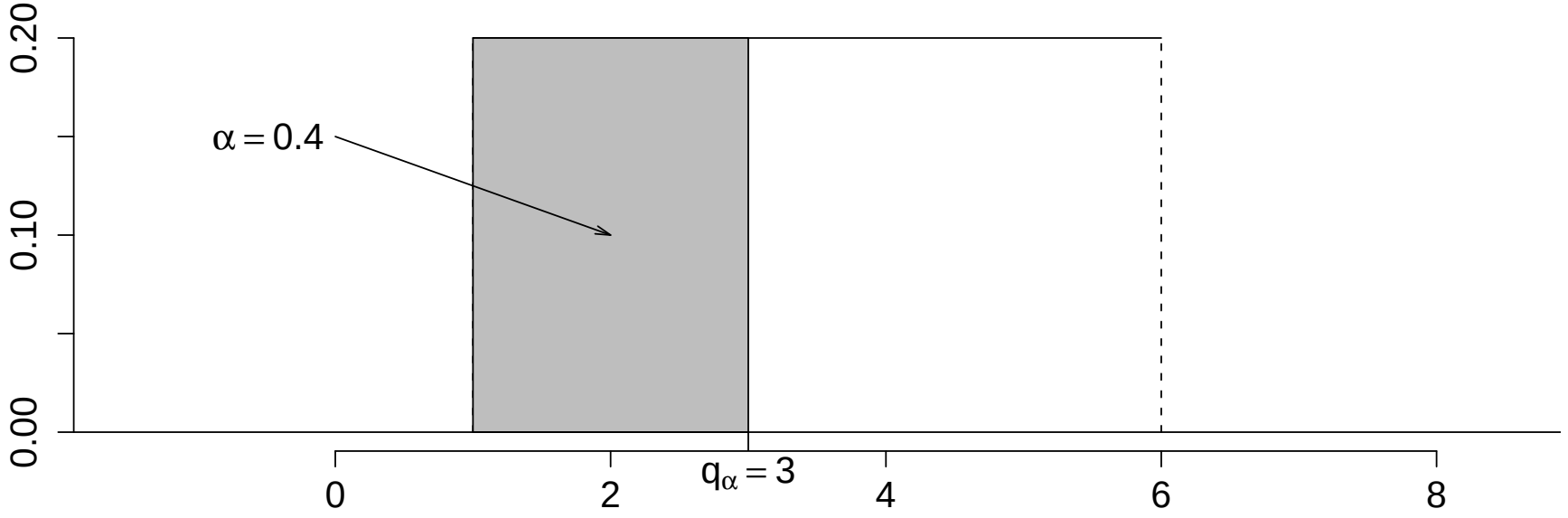
In general (continuous or discrete case) the α-quantile of a cdf $F(x)$ is defined as

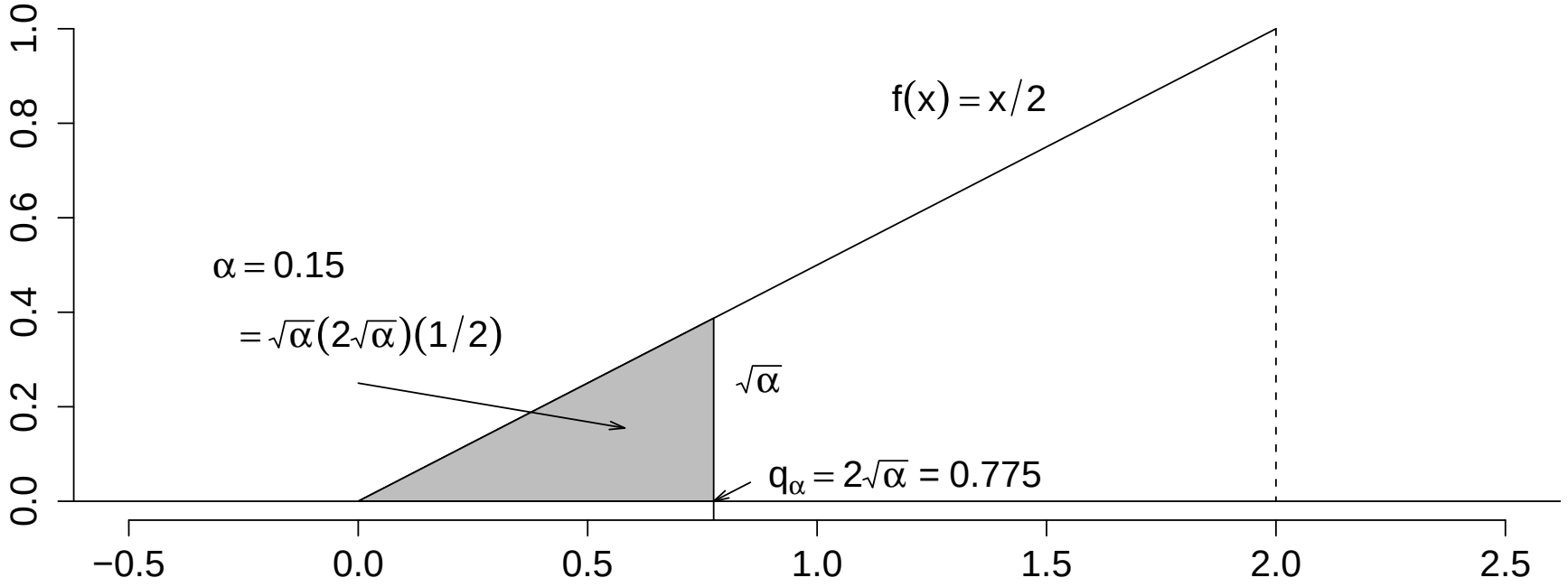
$$q_\alpha = F^{-1}(\alpha) = \text{minimum}\,\{x : F(x) \geq \alpha\}$$

In this definition the quantile is unique (an arbitrary convention).

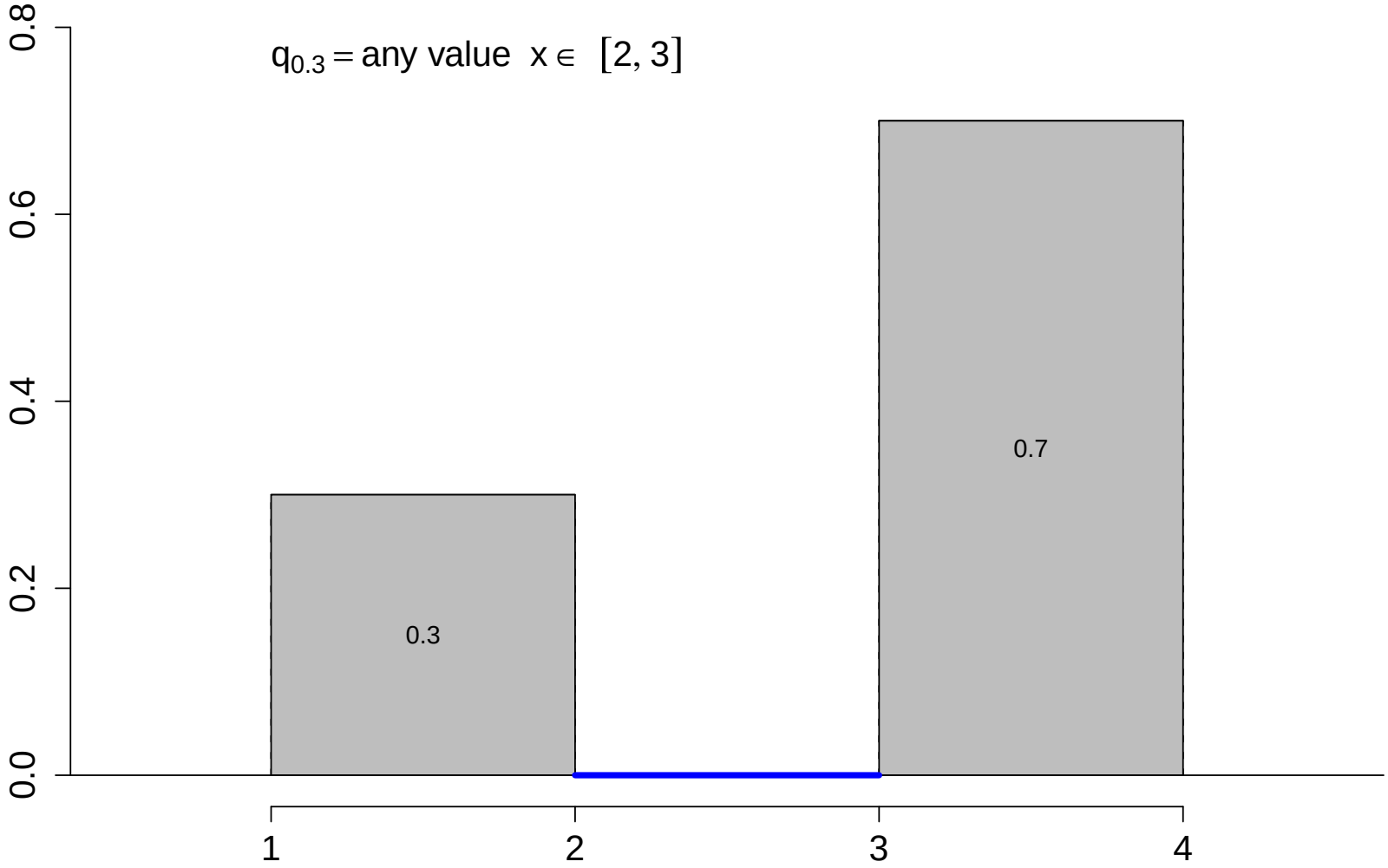
This is the definition that R uses in all its distributions.

Two Quantile Illustrations





Nonunique Quantile Example



$F^{-1}(0.3) =$ left endpoint of the blue interval. Could also choose interval midpoint.

When $\alpha = 0.5$ this midpoint is usually the most popular choice for $q_{0.5} =$ median(X).

Quantiles in R

The .2-quantile of the Uniform[2,4] distribution is obtained as

```
> qunif(.2,2,4)
[1] 2.4
```

The .95-quantile of the standard normal distribution is obtained as

```
> qnorm(.95)
```
or
```
> qnorm(.95,0,1)
[1] 1.644854
```

The median or .5-quantile of a χ^2_{10} distribution is

```
> qchisq(.5,10)
[1] 9.341818
```
note that it is smaller than the mean 10.

The .9-quantile of the t_7 distribution is

```
> qt(.9,7)
[1] 1.414924
```

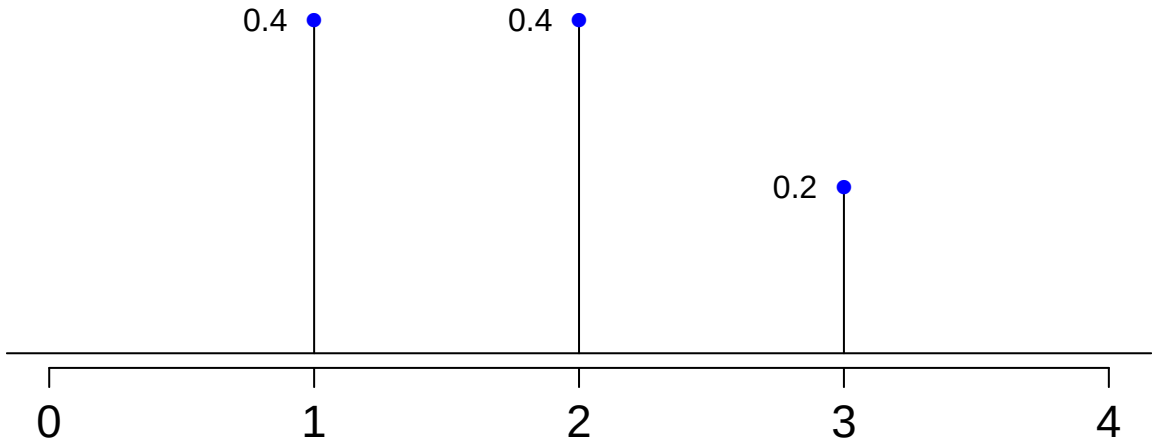
The .8-quantile of the $F_{7,20}$ distribution is

```
> qf(.9,7,20)
[1] 2.039703
```

Quantiles for Discrete Distributions

For discrete distributions it is often not possible to find a q such that

$$P(X < q) = \alpha \qquad \text{and} \qquad P(X > q) = 1 - \alpha \qquad \text{for a given } \alpha \in (0, 1)$$



Consider $\alpha = 0.5$: For $q < 2$ we have $P(X < q) < 0.5 < P(X > q)$

and for $q > 2$ we have $P(X < q) > 0.5 > P(X > q)$.

For $q = 2$ we get: $P(X < q) = 0.4 < 0.5$ and $P(X > q) = 0.2 < 0.5$,

$q_{0.5} = q(X; 0.5) = 2$ is the most appealing choice, since $P(X < q)$ and $P(X > q)$ both cross 0.5 as we move from $q = 2$ to slightly higher or lower, respectively.

Quantile Definition for All Distributions

Definition: For any random variable X and any $\alpha \in (0,1)$ the quantity q is called an α-quantile of X or its distribution whenever

$$P(X < q) \leq \alpha \qquad \text{and} \qquad P(X > q) \leq 1 - \alpha$$

or equivalently (by complementation), whenever

$$P(X < q) \leq \alpha \qquad \text{and} \qquad P(X \leq q) \geq \alpha$$

since

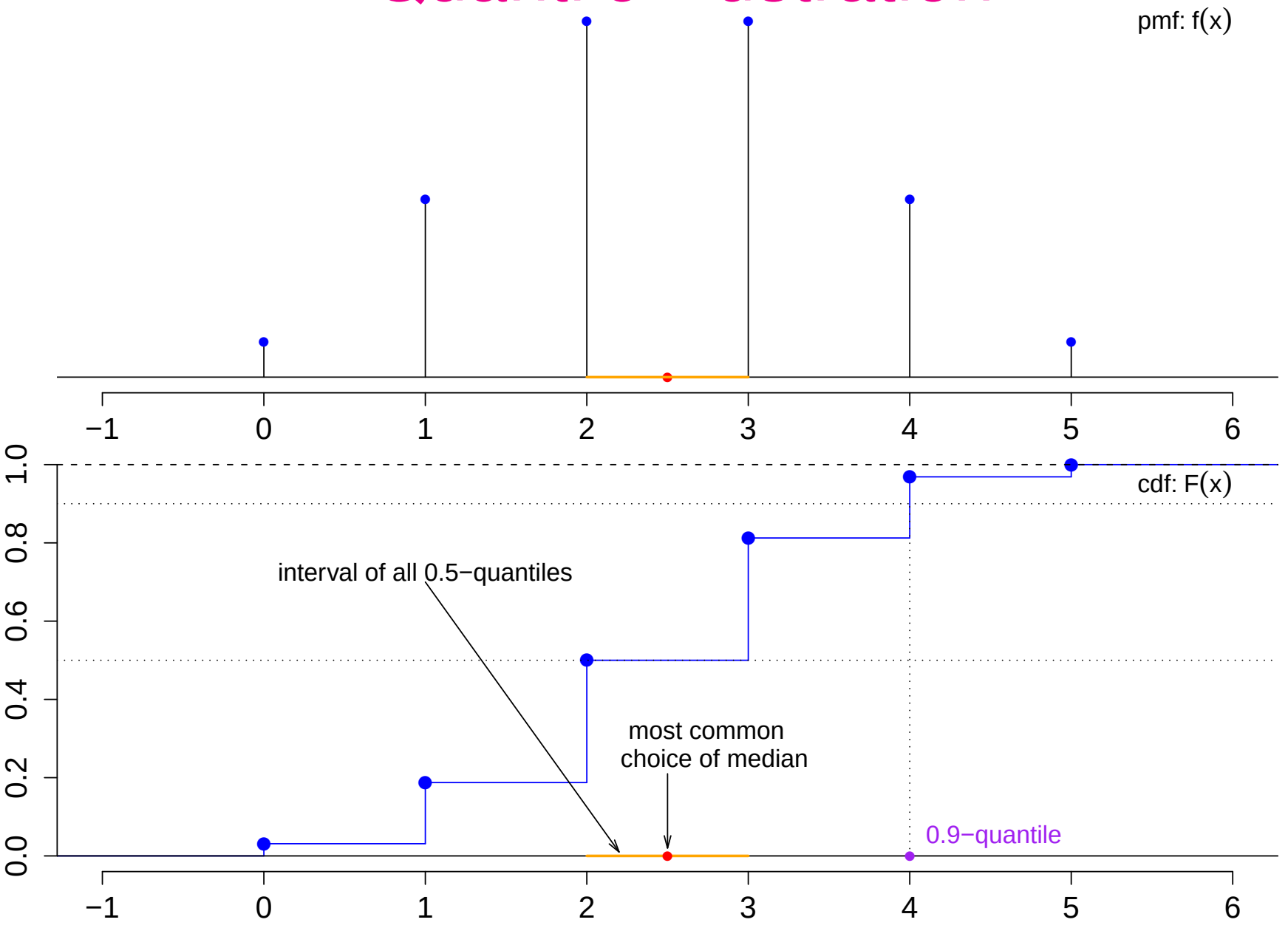
$$P(X > q) \leq 1 - \alpha \implies \alpha \leq 1 - P(X > q) = P(X \leq q)$$

This definition still allows for multiple α-quantiles when $F(x) = \alpha$ for all $x \in [q_1, q_2)$, where $F(x) < \alpha$ for $x < q_1$ and $F(q_2) > \alpha$.

Then this interval consists of all α-quantiles of F.

$F^{-1}(\alpha) = q_1$ always gives you a unique choice for that α-quantile.

Quantile Illustration



Symmetry Revisited

When X has a distribution that is symmetric around θ then the median of this distribution is median$(X) = \theta$. In that case median$(X) = EX$, provided the latter exists.

$$P(X < \theta) = P(X - \theta < 0) = P(X - \theta > 0) = P(X > \theta) = \frac{1 - P(X = \theta)}{2} \leq \frac{1}{2}$$

$$\Longrightarrow \quad \theta = q(X; 0.5) = q_{0.5} = \text{median}(X)$$

For symmetric distributions around θ any ambiguity interval for $q_{0.5}$ is centered on θ.

Thus we prefer the interval midpoint as median, in case of ambiguity.

Outlier Resistance of Median

x	$f(x)$
-1	0.19
0	0.60
1	0.19
10^7	0.02

The median of this distribution is 0.

It stays there, even if we make 10^7 much larger.

Because of this the median is called outlier resistant.

In salary surveys the median salary is often more meaningful than the mean in portraying the midpoint or center of the salary distribution.

For the above distribution we have

$$EX = (-1) \cdot 0.19 + 0 \cdot 0.6 + 1 \cdot 0.19 + 10^7 \cdot 0.02 = 2 \cdot 10^5$$

Interquantile Ranges

As a measure of spread in the X distribution we could take the quantile difference

$$q(X;\alpha_2) - q(X;\alpha_1) = q_{\alpha_2} - q_{\alpha_1} \qquad \text{for any } \alpha_1 < \alpha_2$$

For continuous distributions such an interval $[q_{\alpha_1}, q_{\alpha_2}]$ contains $\alpha_2 - \alpha_1$ of the X-distribution probability.

With that choice of $\alpha_1 < \alpha_2$ we could say that the X distribution is more dispersed than the Y distribution if

$$q(X;\alpha_2) - q(X;\alpha_1) > q(Y;\alpha_2) - q(Y;\alpha_1)$$

since the same amount of probability, $\alpha_2 - \alpha_1$, is spread out over a wider interval for the X distribution than for the Y distribution.

Such ranges are mainly used in the context of continuous distributions.

However, which $\alpha_1 < \alpha_2$ to choose presents too many choices.

Interquartile Range

Statisticians have singled out the interquartile range with $\alpha_1 = 0.25$, $\alpha_2 = 0.75$, i.e,

$$\text{iqr}(X) = q_{0.75} - q_{0.25} = q(X;0.75) - q(X;0.25) = q_3 - q_1$$

The interval $[q_{0.25}, q_{0.75}]$ contains half of the X distribution probability.

The notation $q_3 - q_1$ uses two of a distribution's 3 quartiles q_1, q_2, q_3,

defined for continuous distributions by

$$P(X \le q_1) = 0.25, \quad P(X \le q_2) = 0.5 \quad \text{and} \quad P(X \le q_3) = 0.75$$

Of course, q_2, the second quartile, is just the median.

The 3 quartiles divide a distribution into 4 intervals containing 0.25 probability each.

To some extent the iqr is also outlier resistant.

Alternate Characterizations of EX and median(X)

We could measure dispersion of an X distribution by the expected value of $|X-c|$ or $(X-c)^2$ for some chosen value c, i.e.,

$$E|X-c| \qquad \text{or} \qquad E(X-c)^2$$

Again we have to deal with the choice of c.

Resolve this by choosing c which minimizes the respective dispersion measure.

Theorem: Let X be a random variable with population median q_2 and population mean $\mu = EX$. Then

1) the value of c that minimizes $E|X-c|$ is $c = q_2$
2) the value of c that minimizes $E(X-c)^2$ is $c = \mu$.

The second statement follows from

$$\begin{aligned} E(X-c)^2 &= E(X-\mu+\mu-c)^2 = E(X-\mu)^2 + E[2(X-\mu)(\mu-c)] + E(\mu-c)^2 \\ &= E(X-\mu)^2 + 2(\mu-c)E(X-\mu) + (\mu-c)^2 = E(X-\mu)^2 + (\mu-c)^2 \\ &\geq E(X-\mu)^2 = \sigma^2 \quad \text{an example of the method of least squares} \end{aligned}$$

Property 1) we take without proof.

Indication of Asymmetry

We saw that mean and median coincide for symmetric populations.

When $\mu \neq q_2$ we can conclude that the X distribution is not symmetric.

The wider they are apart the stronger the asymmetry.

Normality and iqr and σ

Let $X = \mu + \sigma Z$ with $Z \sim \mathcal{N}(0,1)$ so that $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\alpha = P(X \le q(X;\alpha)) = P(\mu + \sigma Z \le q(X;\alpha)) = P\left(Z \le \frac{q(X;\alpha) - \mu}{\sigma}\right) = P(Z \le q(Z;\alpha))$$

$$\Longrightarrow \quad q(Z;\alpha) = \frac{q(X;\alpha) - \mu}{\sigma} \quad \text{or} \quad q(X;\alpha) = \mu + \sigma q(Z;\alpha)$$

$$\Longrightarrow \quad \frac{q(X;\alpha_2) - q(X;\alpha_1)}{\sigma} = \frac{q(Z;\alpha_2) - q(Z;\alpha_1)}{1} = q(Z;\alpha_2) - q(Z;\alpha_1)$$

Thus the ratio (interquantile range)$/\sigma$ does not depend on μ and is constant.

In particular, for the interquartile range of normal populations this ratio is

$$\frac{\text{iqr}}{\sigma} = (\texttt{qnorm}(.75) - \texttt{qnorm}(.25)) = 1.348980$$

If we encounter a random variable X for which this is not the case, then X cannot be normally distributed. The stronger the discrepancy from 1.35 the stronger the nonnormality.