A continuous probability distribution: Throwing darts at a dartboard (3.4)

Prof. Tesler

Math 186
Winter 2020

Continuous distributions

Example

- Pick a real number x between 20 and 30 with all real values in $[20, 30]$ equally likely.
- Sample space: $S = [20, 30]$
- Number of outcomes: $|S| = \infty$
- Probability of each outcome: $P(X = x) = \frac{1}{\infty} = 0$
- Yet, $P(X \leqslant 21.5) = 15\%$

Continuous distributions

- The *sample space* S is often a subset of $\mathbb{R}^n$.
  We'll do the 1-dimensional case $S \subset \mathbb{R}$.
- The *probability density function (pdf)* $f_X(x)$ is defined differently than the discrete case:
  - $f_X(x)$ is a real-valued function on S with $f_X(x) \geqslant 0$ for all $x \in S$.
  - $\int_S f_X(x)\,dx = 1 \quad$ (vs. $\sum_{x \in S} p_X(x) = 1$ for discrete)
  - The probability of event $A \subset S$ is $P(A) = \int_A f_X(x)\,dx \quad$ (vs. $\sum_{x \in A} p_X(x)$).
  - In n dimensions, use n-dimensional integrals instead.

Uniform distribution

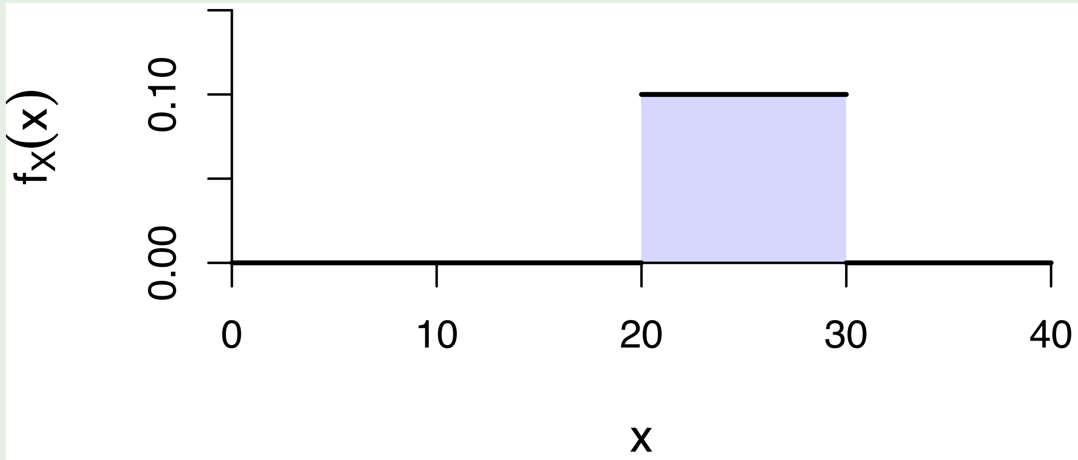
- Let $a < b$ be real numbers.
- The *Uniform Distribution* on $[a, b]$ is that all numbers in $[a, b]$ are "equally likely."
- More precisely, $f_X(x) = \begin{cases} \frac{1}{b-a} & \text{if } a \leqslant x \leqslant b; \\ 0 & \text{otherwise.} \end{cases}$

Uniform distribution (real case)

The uniform distribution on $[20, 30]$

We could regard the sample space as $[20, 30]$, or as all reals.

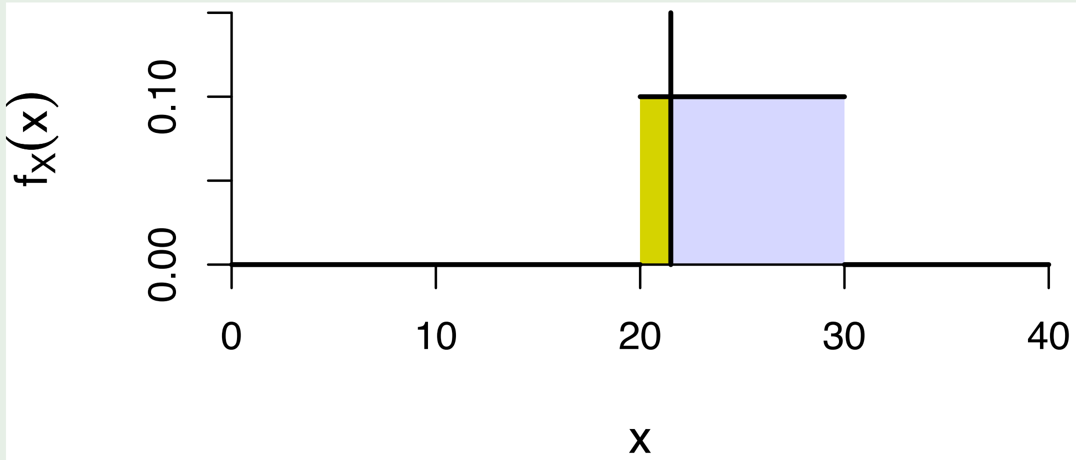
$$f_X(x) = \begin{cases} 1/10 & \text{for } 20 \leqslant x \leqslant 30; \\ 0 & \text{otherwise.} \end{cases}$$



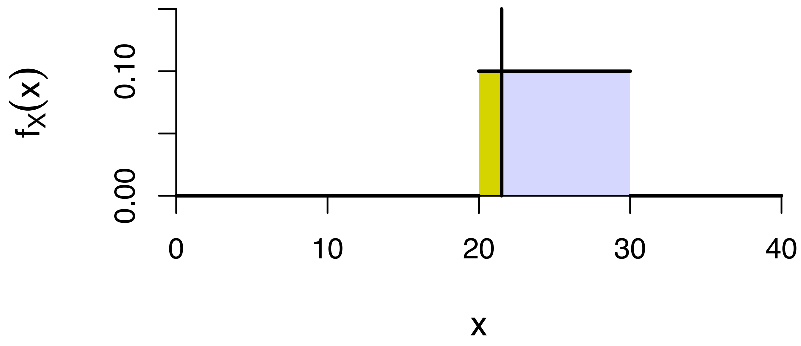
$$P(X \leqslant 21.5) = \int_{-\infty}^{20} 0\,dx + \int_{20}^{21.5} \frac{1}{10}dx = 0 + \left.\frac{x}{10}\right|_{20}^{21.5}$$

$$= \frac{21.5 - 20}{10}$$

$$= .15 = 15\%$$



Probability of a point, in the continuous case



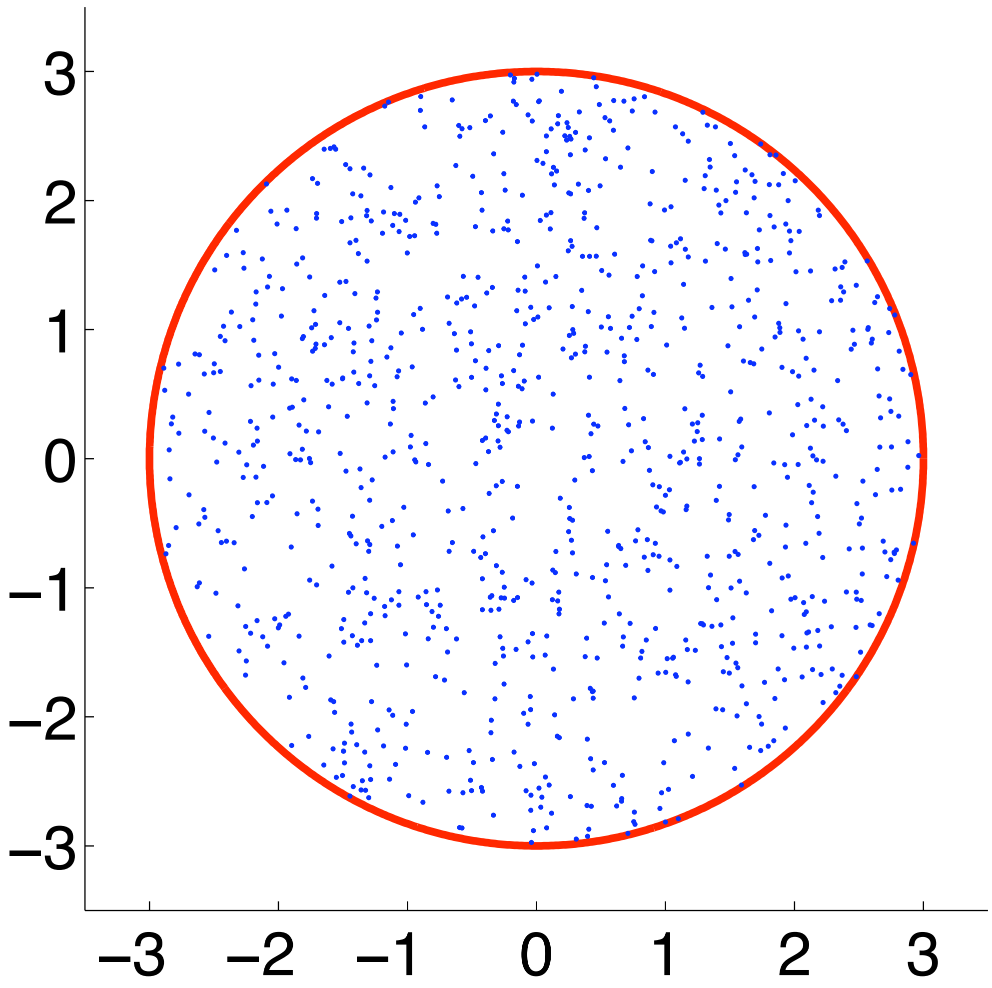
Probability of a point

- For any continuous random variable, the probability of a point, b, is
  $$P(X = b) = \int_b^b f_X(x)\,dx = \text{area of line segment} = 0.$$
- The *probability density* $f_X(b)$ may be nonzero, but integrating over a single point gives *probability* $= 0$.

$P(X \leqslant b) = P(X < b) + P(X = b)$

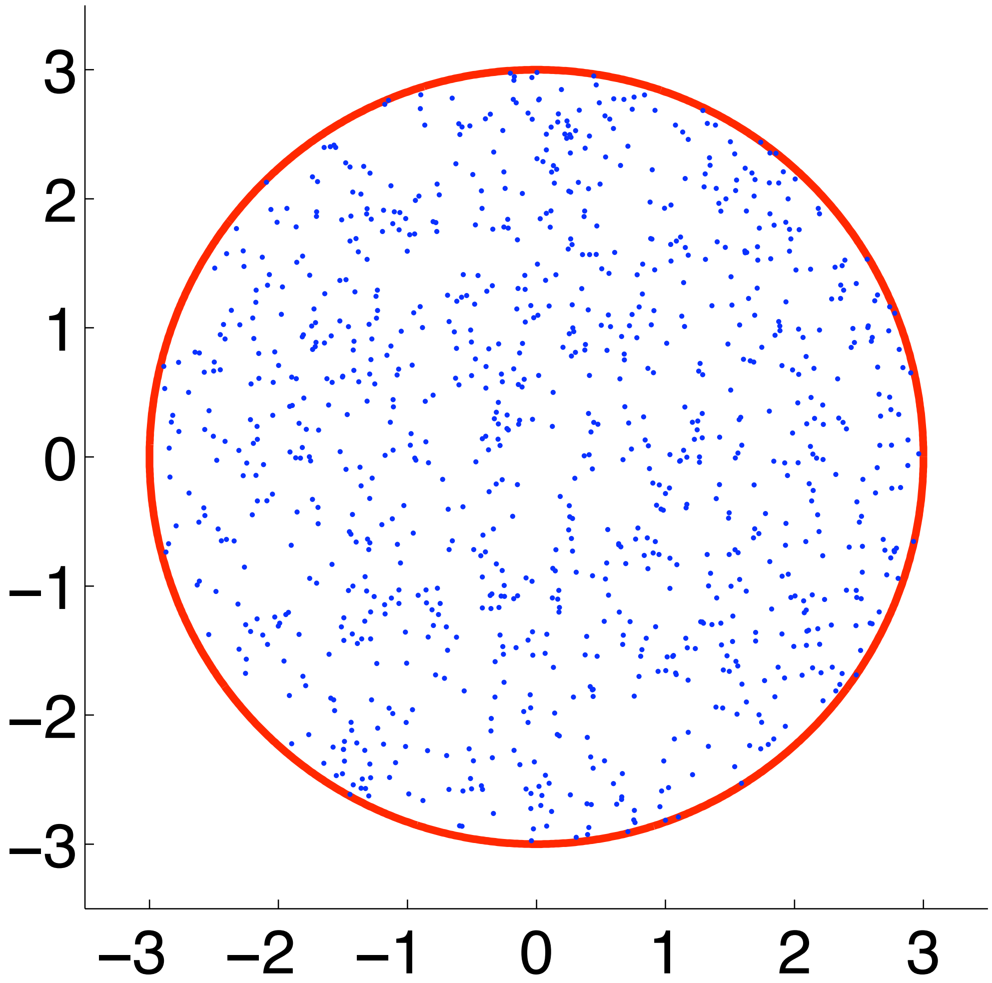
- **Continuous case:** $P(X = b) = 0$, so $P(X \leqslant b) = P(X < b)$.
  Similarly, $P(a \leqslant X) = P(a < X)$.
- **Discrete case:** $P(X = b) = p_X(b) \geqslant 0$. If nonzero, then
  $$P(X \leqslant b) \neq P(X < b).$$

Dartboard



- A dart is repeatedly thrown at a dartboard.
- **Shape:**
  Circle of radius 3
  centered at the origin.
- Assume all points on the board are hit with equal ("uniform") probability (and ignore the darts that miss).

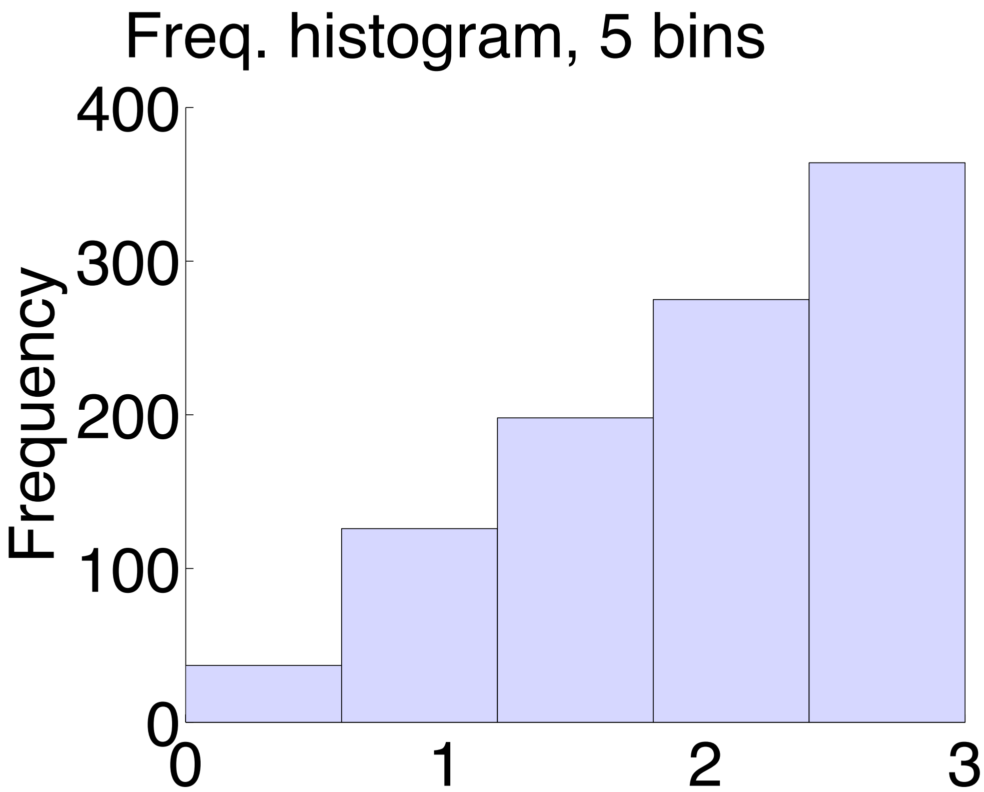
Data



i	x	y	r
1	1.575	1.022	1.878
2	-1.640	1.265	2.071
3	-1.625	0.607	1.734
4	-1.143	-1.947	2.257
5	-1.054	-0.822	1.337
...	...	...	...
999	1.747	0.850	1.943
1000	1.519	-1.429	2.086

- Hits $i = 1, 2, \ldots, 1000$.
- Coordinates (x_i, y_i).
- Distance to center $r_i = \sqrt{{x_i}^2 + {y_i}^2}$.
- *What is the distribution of r?*

Histograms – Frequency Histogram

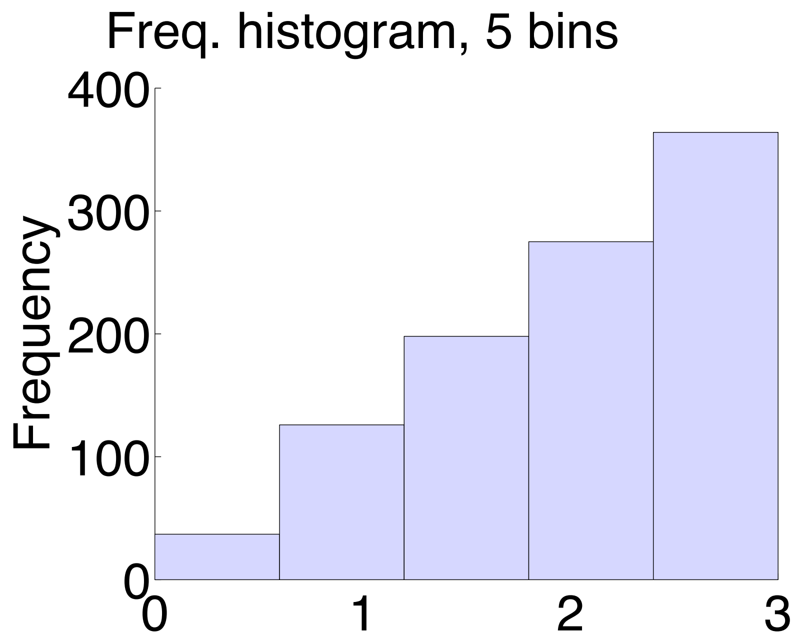


5 bins

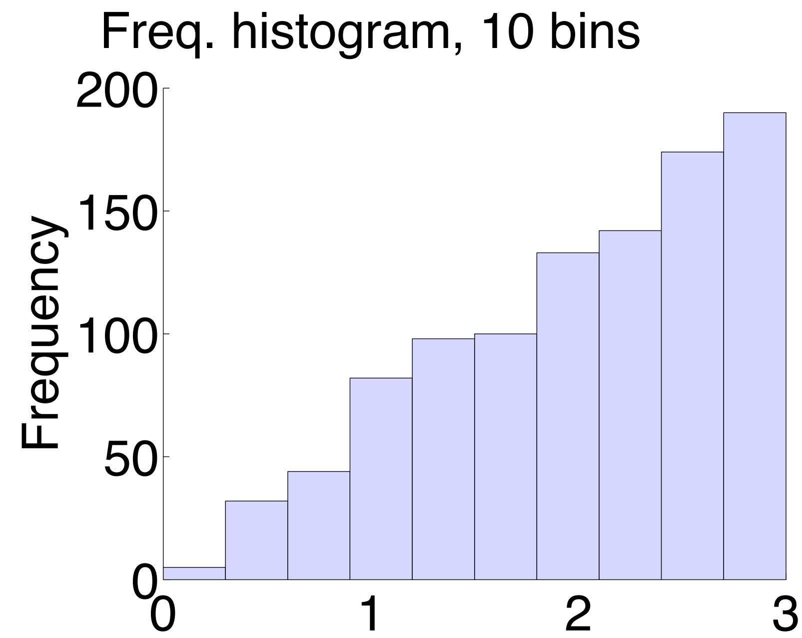
- $w =$ bin width $= 3/5 = 0.6$
- **Bins (x-axis):**
  $0 \leqslant r < 0.6$ has n_1=37 points
  $0.6 \leqslant r < 1.2$ has $n_2 = 126$
  $1.2 \leqslant r < 1.8$ has $n_3 = 198$
  $1.8 \leqslant r < 2.4$ has $n_4 = 275$
  $2.4 \leqslant r \leqslant 3.0$ has $n_5 = 364$
  Total $n = n_1 + \cdots + n_5 = 1000$

Notation

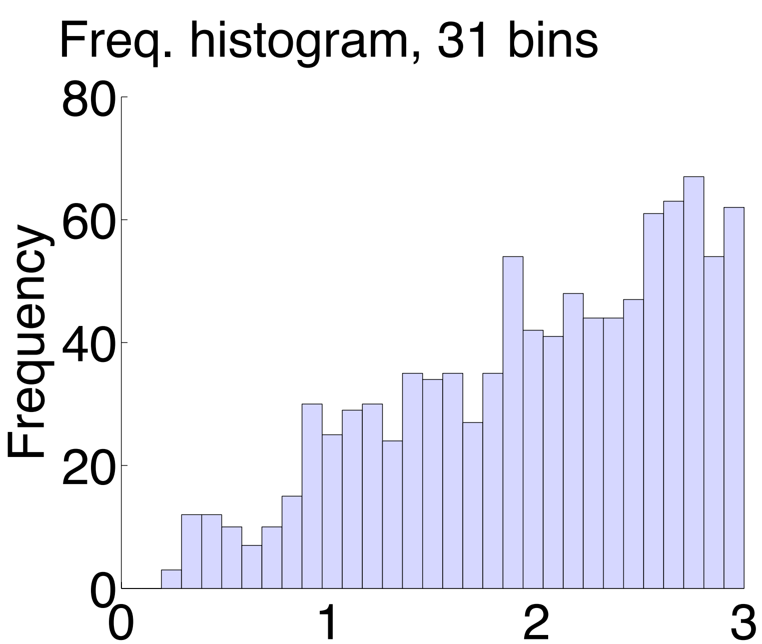
- $n =$ # points $= 1000$
- $w =$ bin width $= 3/$(# bins)
- $n_j =$ # points in bin j
- **y-axis:** Set bar height = n_j
- **Area of bin j:**
  width $\times$ height $= w \cdot n_j$
- **Area of histogram:**
  $A = w(n_1 + n_2 + \cdots) = w \cdot n$



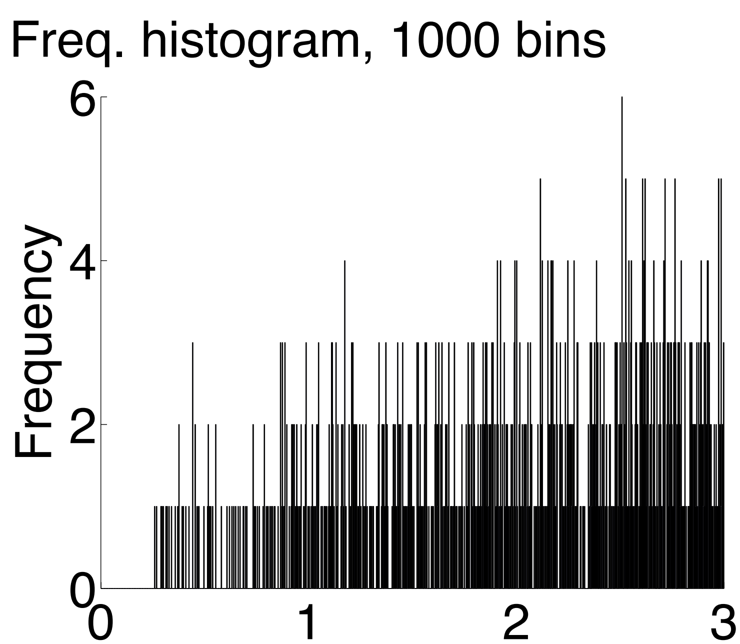
$w = 3/5$, $A = (3/5)(1000) = 600$



$w = 3/10$, $A = (3/10)(1000) = 300$

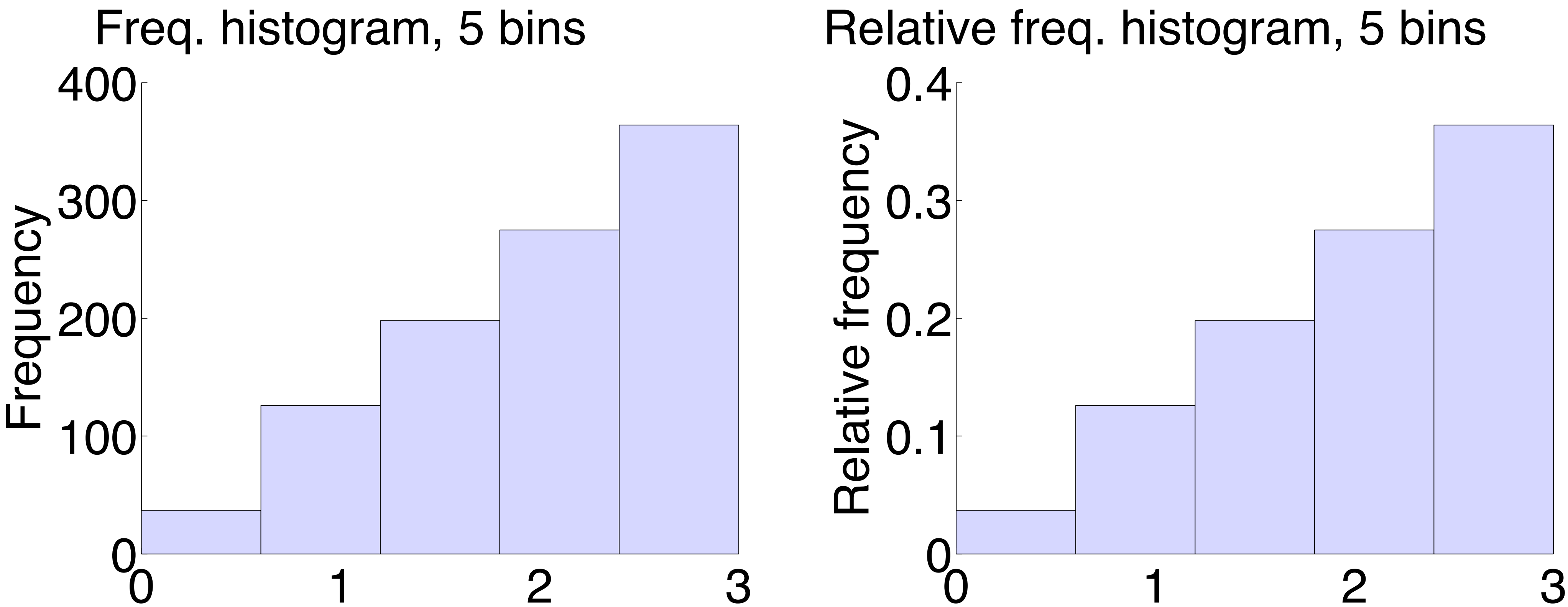


$w = 3/31$, $A = (3/31)(1000) \approx 96.77$

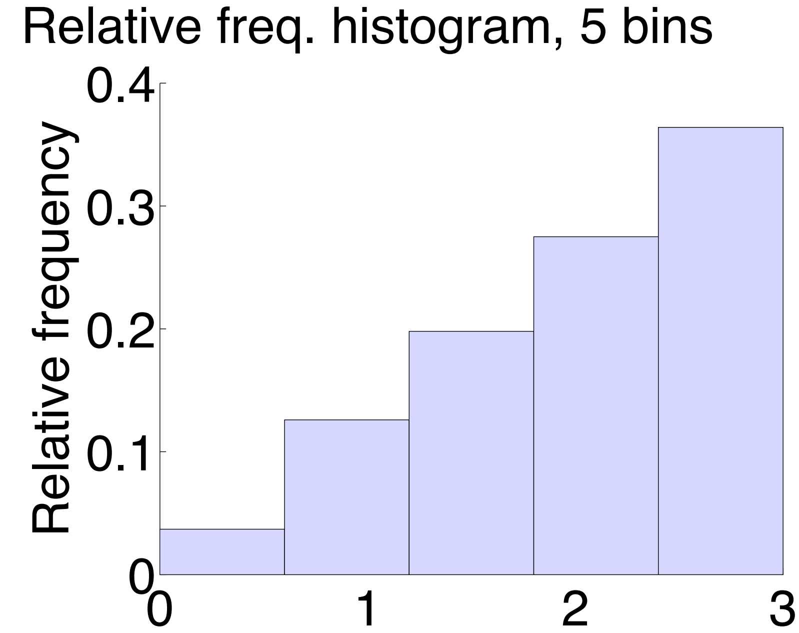


$w = 3/1000$, $A = (3/1000)(1000) = 3$

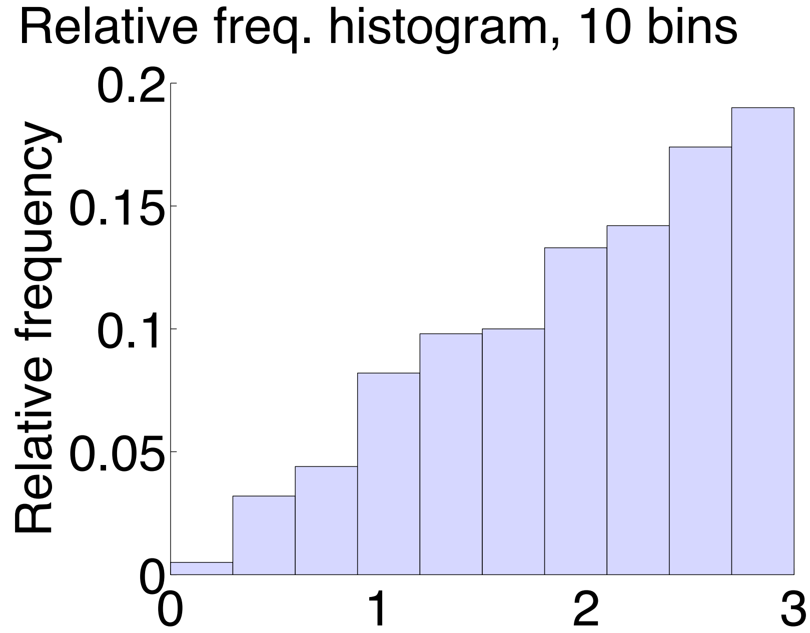
Relative frequency histogram



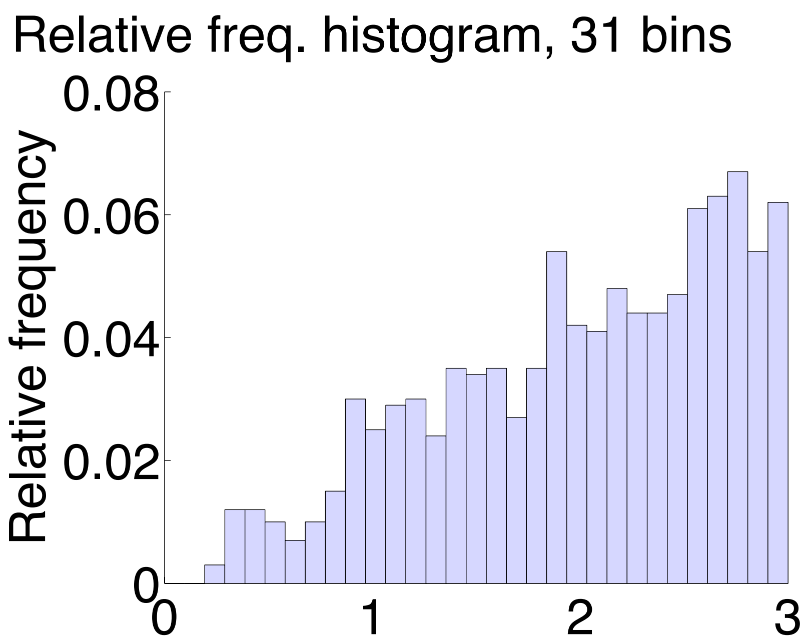
- **Relative frequency:** the fraction of points in bin j is n_j/n.
- Plot a bar of height n_j/n instead of height n_j.
- Bar j area = $w \cdot n_j/n$
- Total area = $w \cdot (n_1 + n_2 + \cdots)/n = w \cdot n/n = w$.
- Graphs look the same, just the y-axis scale changes.



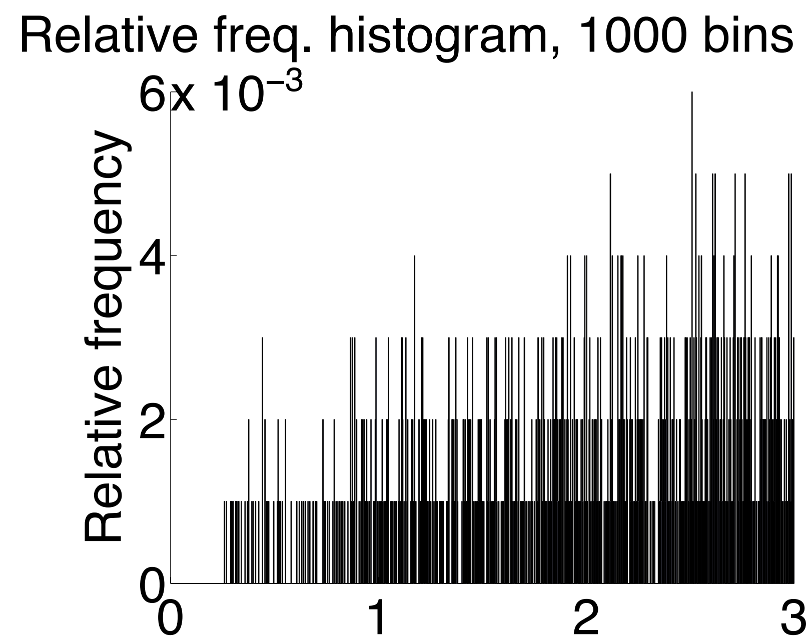
$w = 3/5$, $A = 3/5$



$w = 3/10$, $A = 3/10$

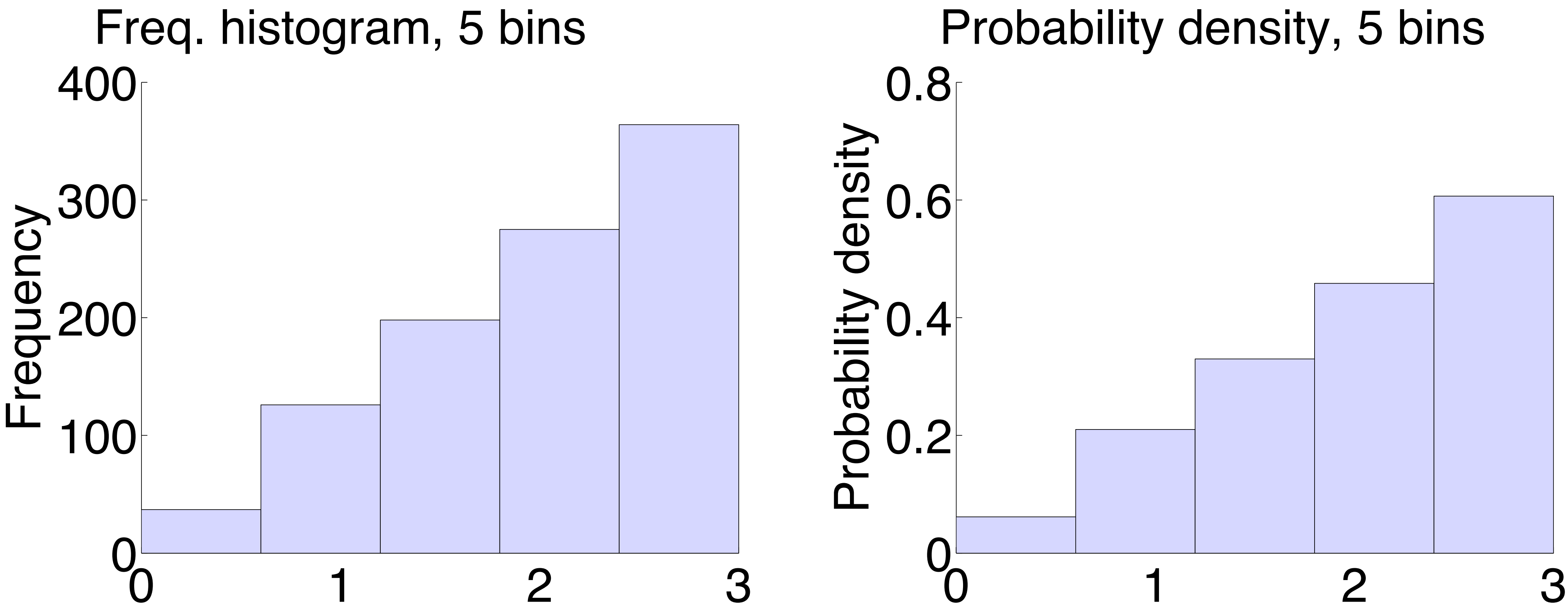


$w = 3/31$, $A = 3/31$

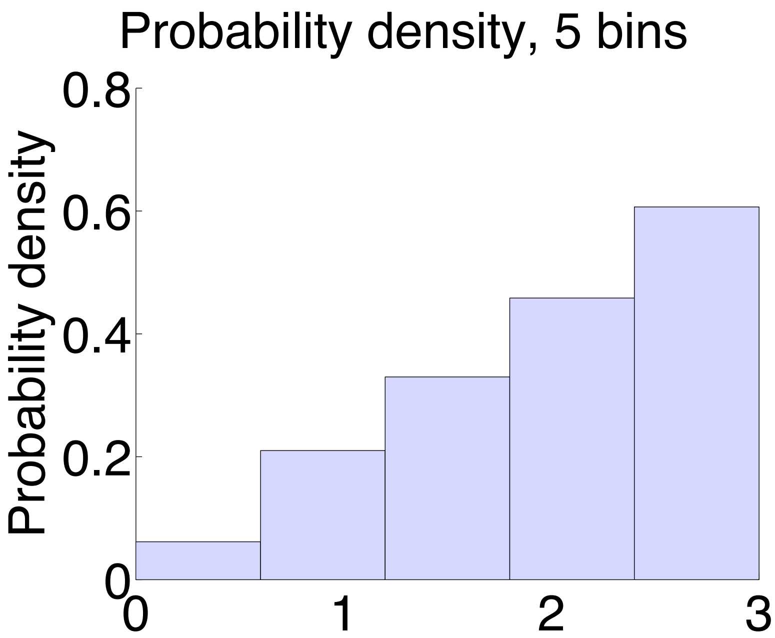


$w = 3/1000$, $A = 3/1000$

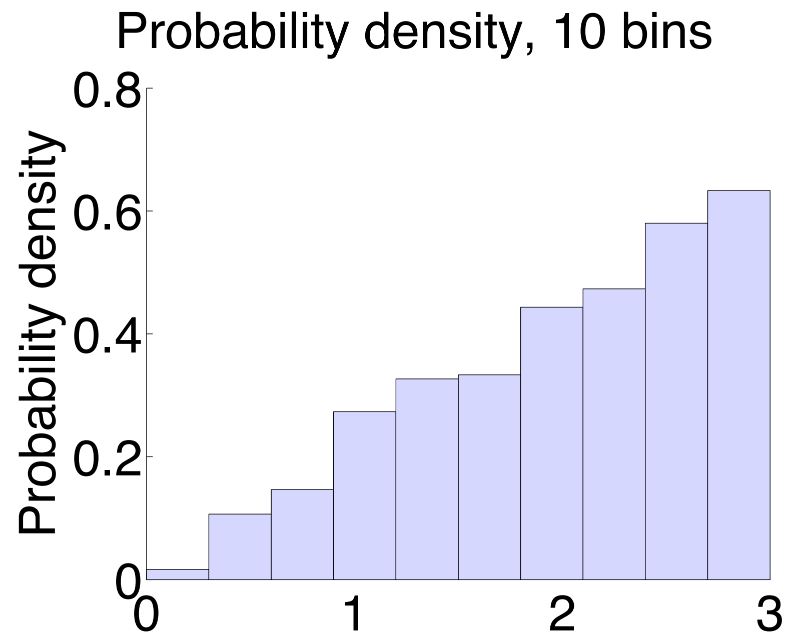
Probability density histogram



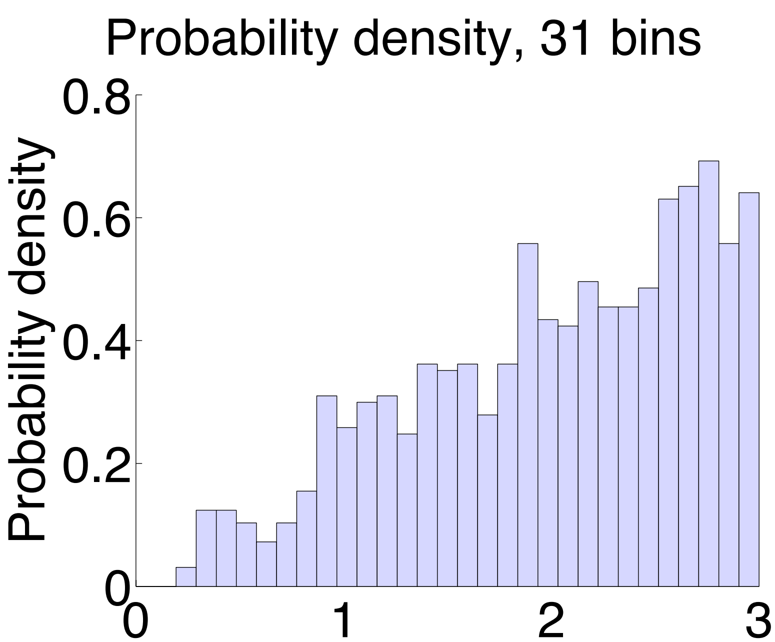
- **Probability density per unit x:** the fraction of points in bin j is n_j/n, and bin j has width w, giving density $n_j/(nw)$.
- Plot a bar of height $n_j/(nw)$
- Bar j area = $w \cdot n_j/(nw) = n_j/n$
- Total area = $(n_1 + n_2 + \cdots)/n = n/n = 1$.
- Graphs look the same, just the y-axis scale changes.



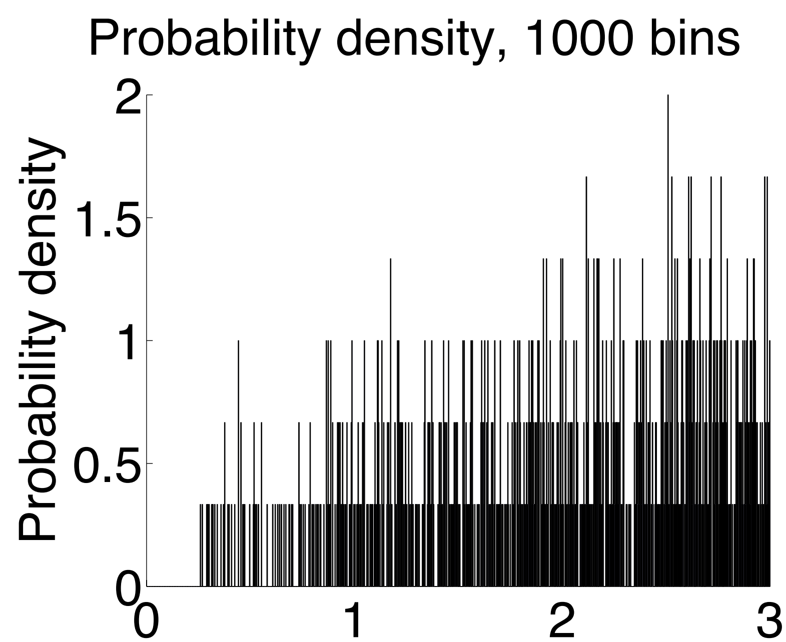
$w = 3/5,\ A = 1$



$w = 3/10,\ A = 1$



$w = 3/31,\ A = 1$



$w = 3/1000,\ A = 1$

How many bins?

How many bins to use?

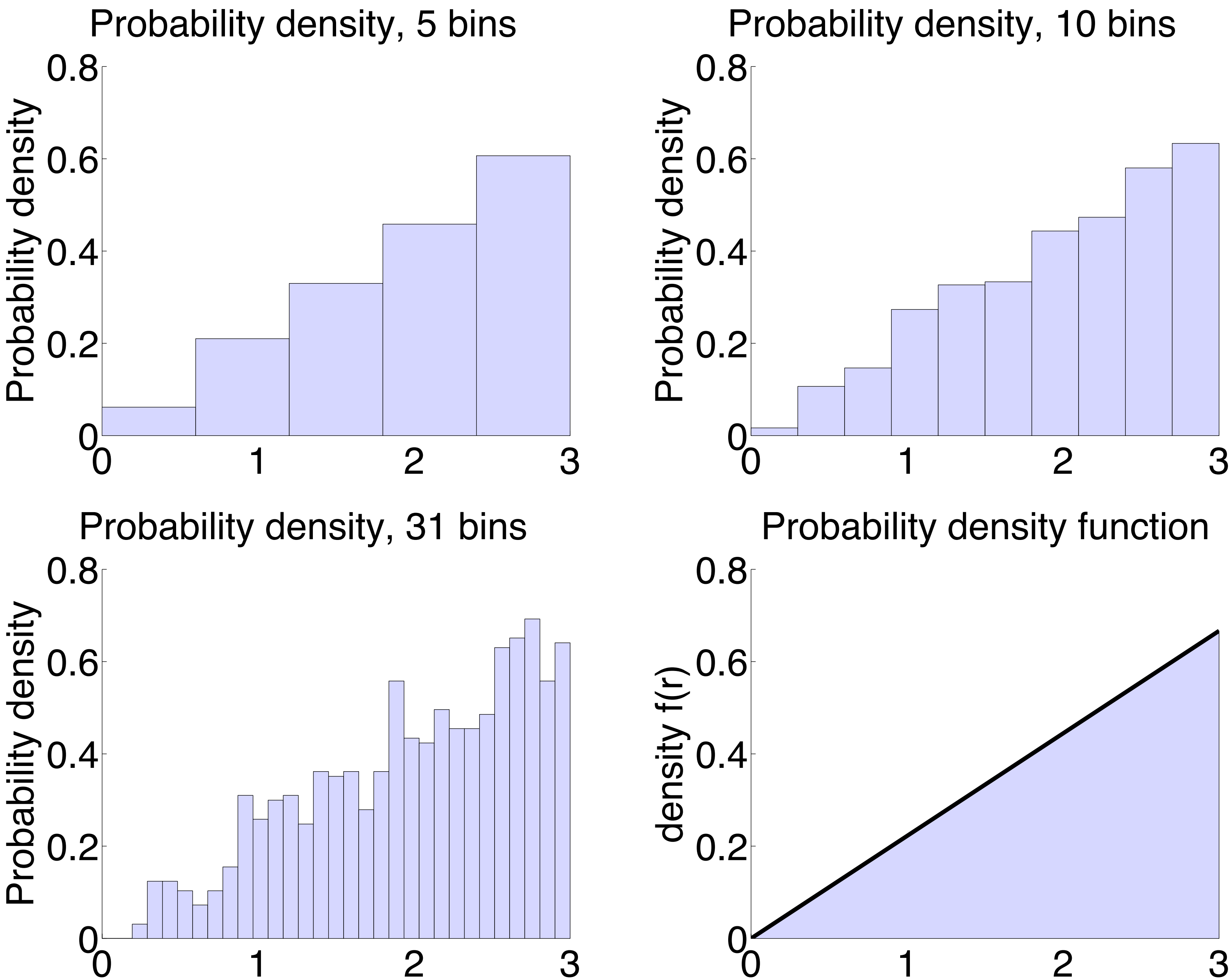
- For $n = 1000$ points, 5, 10, 31 bins all looked reasonable, while 1000 bins did not (too many empty or overfilled bins).
- Usually use a small fixed number of bins, much smaller than the number of points.
- In the discrete case, sometimes it's a concern to pick bin boundaries so that the points don't hit the boundaries.

Effect on y-axis of changing number of bins

- **Frequency and relative frequency histograms:** Increasing the number of bins cuts the y-axis proportionately.
- **Probability density histogram:** Increasing the number of bins keeps the y-axis stable, as long as the number of bins is much smaller than the number of points.

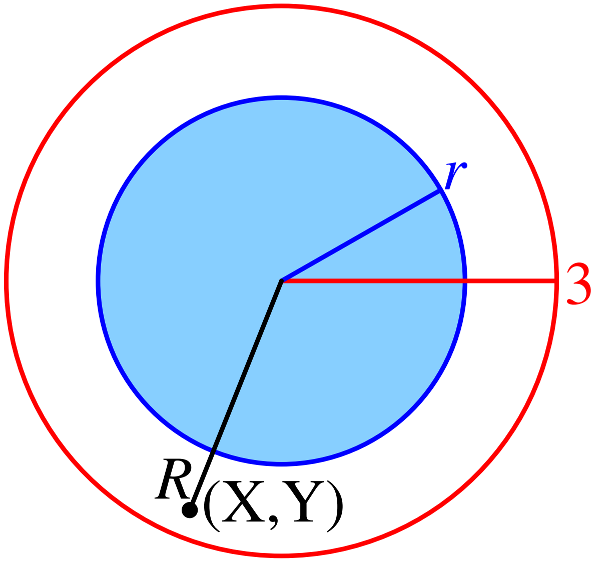
Limit as # points and bins $\to \infty$

$n \to \infty$ and number of bins $\to \infty$ but slower (e.g., $\sqrt{n}$ bins)



Probability density function (PDF) of a continuous random variable

- Let X, Y be random variables for the coordinates of a random point in the circle and $R = \sqrt{X^2 + Y^2}$.



- For each r between 0 and 3,

$$P(R \leqslant r) = \text{Area of circle of radius } r \text{ (centered at origin)}$$
$$\div \text{ Area of whole circle}$$
$$= (\pi r^2)/(\pi 3^2) = r^2/9$$

Also, $P(R \leqslant r) = 0$ if $r < 0$ and $P(R \leqslant r) = 1$ if $r > 3$.

Probability density function (PDF) of a continuous random variable

- Together:

$$P(R \leqslant r) = \begin{cases} 0 & \text{if } r < 0; \\ r^2/9 & \text{if } 0 \leqslant r \leqslant 3; \\ 1 & \text{if } r \geqslant 3. \end{cases}$$

- But the area up to r in the probability density histogram is

$$P(R \leqslant r) = \int_0^r f(t)\, dt$$

so for $0 \leqslant r \leqslant 3$,

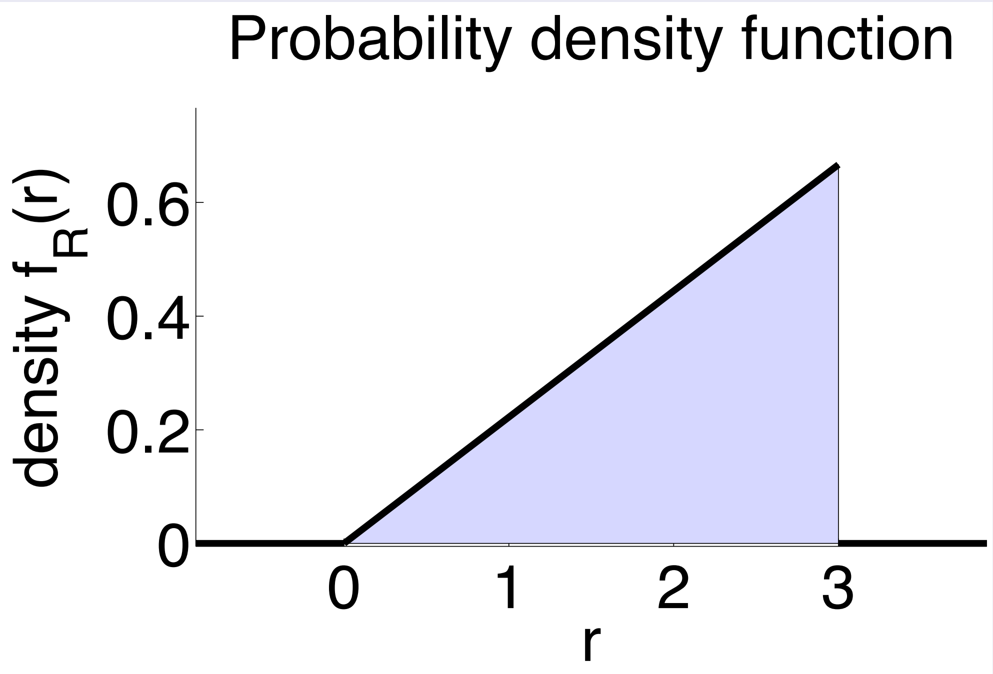
$$f(r) = \frac{d}{dr} P(R \leqslant r) = \frac{d}{dr}\frac{r^2}{9} = \frac{2r}{9}$$

- If $r < 0$ then $f(r) = \frac{d}{dr} 0 = 0$; if $r > 3$ then $f(r) = \frac{d}{dr}(1) = 0$

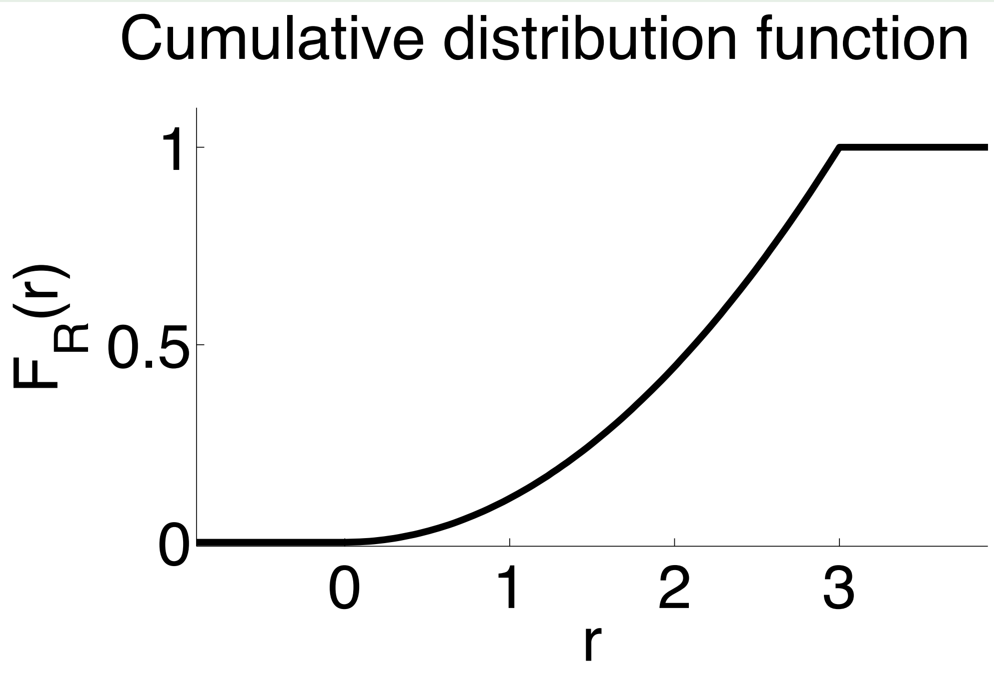
Cumulative distribution function (cdf), continuous case

- The *Cumulative Distribution Function (cdf)* of a random variable is
$$F_X(x) = P(X \leqslant x)$$
- We computed that the cdf of R is
$$F_R(r) = P(R \leqslant r) = \begin{cases} 0 & \text{if } r < 0; \\ r^2/9 & \text{if } 0 \leqslant r \leqslant 3; \\ 1 & \text{if } r \geqslant 3. \end{cases}$$
and then we differentiated it to get the pdf
$$f_R(r) = \begin{cases} 2r/9 & \text{if } 0 \leqslant r < 3; \\ 0 & \text{if } r \leqslant 0 \text{ or } r > 3 \end{cases}$$

PDF vs. CDF



- $f_R(r) = \begin{cases} 2r/9 & \text{if } 0 \leqslant r < 3; \\ 0 & \text{if } r \leqslant 0 \text{ or } r > 3 \end{cases}$
  It's discontinuous at $r = 3$.
- **PDF is derivative of CDF:**
  $f_R(r) = F_R{}'(r)$



- $F_R(r) = \begin{cases} 0 & \text{if } r < 0; \\ r^2/9 & \text{if } 0 \leqslant r \leqslant 3; \\ 1 & \text{if } r \geqslant 3. \end{cases}$
- **CDF is integral of PDF:**
  $F_R(r) = \int_{-\infty}^{r} f_R(t)\, dt$

Probability of an interval

Compute $P(-1 \leqslant R \leqslant 2)$ from the PDF and also from the CDF

Computation from the PDF

$$
\begin{aligned}
P(-1 \leqslant R \leqslant 2) &= \int_{-1}^{2} f_R(r)\,dr = \int_{-1}^{0} f_R(r)\,dr + \int_{0}^{2} f_R(r)\,dr \\
&= \int_{-1}^{0} 0\,dr + \int_{0}^{2} \frac{2r}{9}\,dr \\
&= 0 + \left(\left. \frac{r^2}{9} \right|_{r=0}^{2} \right) = \frac{2^2 - 0^2}{9} = \boxed{\mathbf{\frac{4}{9}}}
\end{aligned}
$$

Computation from the CDF

$$
\begin{aligned}
P(-1 \leqslant R \leqslant 2) &= P(-1^- < R \leqslant 2) \\
&= F_R(2) - F_R(-1^-) = \frac{2^2}{9} - 0 = \boxed{\mathbf{\frac{4}{9}}}
\end{aligned}
$$

Cumulative distribution function (cdf)

For any random variable X, the *Cumulative Distribution Function (cdf)* is $F_X(x) = P(X \leqslant x)$

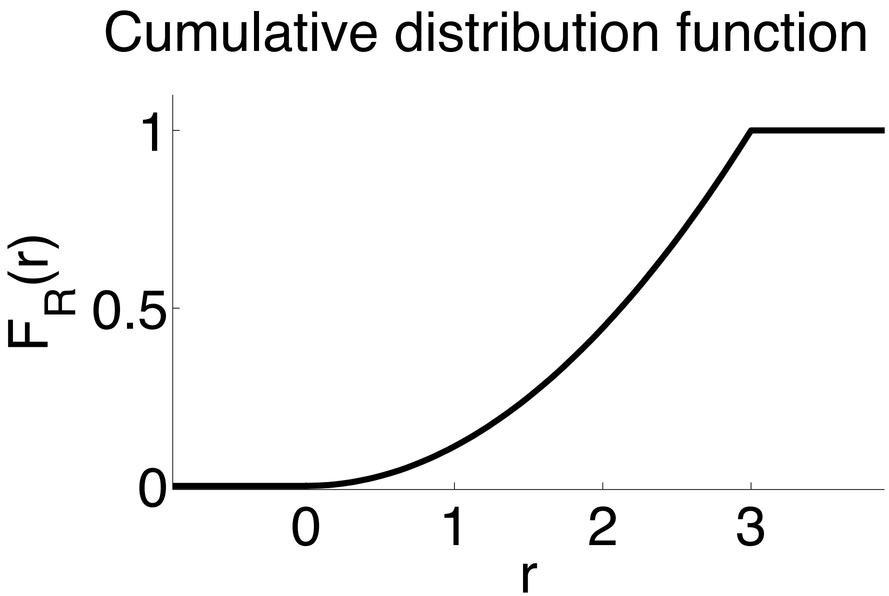
Continuous case

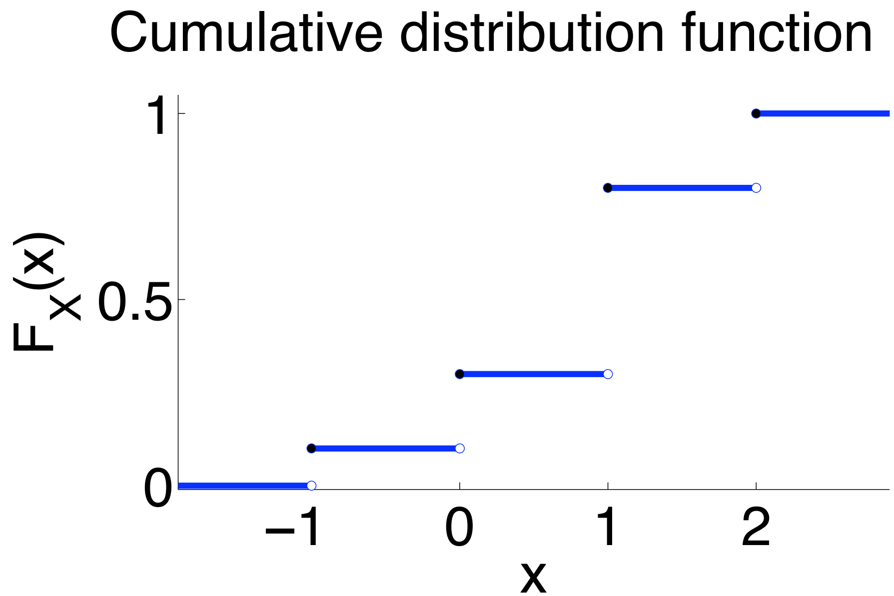
- $F_X(x) = \int_{-\infty}^{x} f_X(t)\,dt$
- Weakly increasing.
- Varies smoothly from 0 to 1 as x varies from $-\infty$ to ∞.
- To get the pdf from the cdf, use $f_X(x) = F_X{}'(x)$.

Discrete case

- $F_X(x) = \sum_{t \leqslant x} p_X(t)$
- Weakly increasing.
- Stair-steps from 0 to 1 as x goes from $-\infty$ to ∞.
- The cdf jumps where $p_X(x) \neq 0$ and is constant in-between.
- To get the pdf from the cdf, use $p_X(x) = F_X(x) - F_X(x^-)$ (which is positive at the jumps, 0 otherwise).

Continuous vs. discrete random variables





In a continuous distribution:

- The probability of an individual point is 0: $P(R = r) = 0$.
  So, $P(R \leqslant r) = P(R < r)$, i.e., $F_R(r) = F_R(r^-)$.
- The CDF is continuous.
  (In a discrete distribution, the CDF is discontinuous due to jumps at the points with nonzero probability.)
- $$P(a < R < b) = P(a \leqslant R < b) = P(a < R \leqslant b) = P(a \leqslant R \leqslant b)$$
  $$= F_R(b) - F_R(a)$$

CDF, percentiles, and median

- The k^{th} *percentile* of a distribution X is the point x where k% of the probability is up to that point:

$$F_X(x) = P(X \leqslant x) = k\% = k/100$$

- **Dartboard:** $F_R(r) = P(R \leqslant r) = r^2/9 \quad$ (for $0 \leqslant r \leqslant 3$)
- $r^2/9 = (k/100) \quad \Rightarrow \quad r = \sqrt{9(k/100)}$
- 75$^{\text{th}}$ percentile: $r = \sqrt{9(.75)} \approx 2.60$
- Median (50$^{\text{th}}$ percentile): $r = \sqrt{9(.50)} \approx 2.12$
- 0$^{\text{th}}$ and 100$^{\text{th}}$ percentiles:
    - $r = 0$ and $r = 3$ on the range $0 \leqslant r \leqslant 3$.
    - Not uniquely defined if r ranges over all real numbers, since
      $F_R(r) = 0$ for all $r \leqslant 0 \quad$ and $\quad F_R(r) = 1$ for all $r \geqslant 3$.