# Random Variables in Multiple Dimensions

Tuesday · joint, marginal, conditional, and independent

# Random Vectors

- We've dealt mostly with a single random variable X that maps the sample/state space $\mathcal{S}$ into a number, X: $\mathcal{S} \rightarrow \mathbb{R}$
- An n-dimensional vector of random variables is a mapping X: $\mathcal{S} \rightarrow \mathbb{R} \times \mathbb{R} \times ... \times \mathbb{R} = \mathbb{R}^L$
- This is more complex than one dimension: **We want to track how random variables move together**
- We will look at the **joint** and **marginal** distributions, and then a different way to think of **conditional probability** and **independence** in these terms

# 1. Joint, Marginal, and Conditional

# Joint, Marginal, Conditional Probability Distributions

- **Joint**: the probability of both
  $m_{X,Y}(x, y) = \text{pr}[X = x, Y = y]$ "P[X AND Y]"
- **Marginal**: "integrate X out"
  $m_Y(y) = \sum_x m_{X,Y}(x, y)$ "P[Y]"
- **Conditional**: select a row, then renormalize "P[Y|X]"
  $\text{pr}[Y = y|X = x] = \frac{m_{X,Y}(x, y)}{m_X(x)} = \frac{\text{joint}}{\text{marginal}}$
- Empirically, every m is a proportion of rows in the data

| | **Y = 0** | **Y = 1** | **Marginal** |
|---|---|---|---|
| **X = a** | joint | joint | row sum |
| **X = b** | joint | joint | row sum |
| **Marginal** | col sum | col sum | 1 |

Conditional on X = a: divide row a by its row sum

You can get this table by using pd.crosstab on X and Y

$$\mathit{CONDITIONAL} = \frac{\mathit{JOINT}}{\mathit{MARGINAL}}$$

# Worked Example: Calculate the joint & marginal distributions

| Person | Smoker | Sick |
|---|---|---|
| 1 | Yes | Yes |
| 2 | Yes | Yes |
| 3 | Yes | No |
| 4 | No | No |
| 5 | No | Yes |

X = Smoker, Y = Sick

- **Step 1: count** each (Smoker, Sick) combination. Add row and column totals

| | Sick | Not sick | Total |
|---|---|---|---|
| **Smoker** | 2 | 1 | 3 |
| **Non-smoker** | 1 | 1 | 2 |
| **Total** | 3 | 2 | 5 |

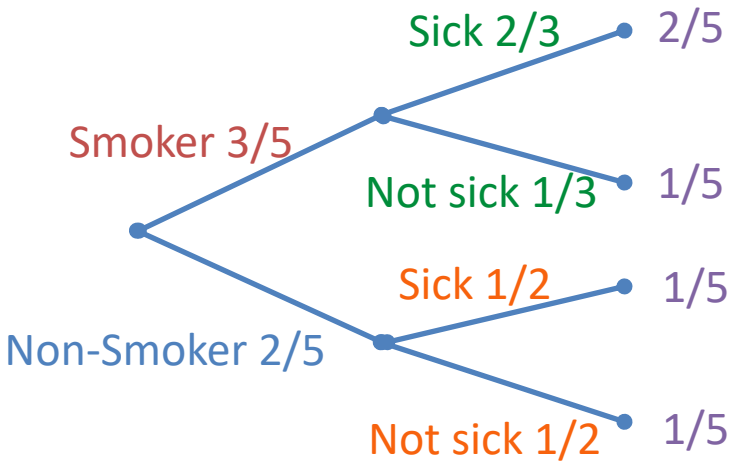
- **Step 2: joint** = count / 5. The **margins** are the row and column sums

| | Sick | Not sick | Marginal |
|---|---|---|---|
| **Smoker** | 2/5 | 1/5 | 3/5 |
| **Non-smoker** | 1/5 | 1/5 | 2/5 |
| **Marginal** | 3/5 | 2/5 | 1 |

The whole table sums to 1

- **Joint**: pr[Smoker, Sick] = 2/5. Two of the five people are both
- **Marginal**: pr[Sick] = 2/5 + 1/5 = 3/5. Sum down the Sick column, ignoring smoking
- In general, marginal: To describe just one of the two variables, sum over the other

# Worked Example: Conditional and Independence




| Person | Smoker | Sick |
|---|---|---|
| 1 | Yes | Yes |
| 2 | Yes | Yes |
| 3 | Yes | No |
| 4 | No | No |
| 5 | No | Yes |

| | Sick | Not sick | Marginal |
|---|---|---|---|
| **Smoker** | 2/5 | 1/5 | 3/5 |
| **Non-smoker** | 1/5 | 1/5 | 2/5 |
| **Marginal** | 3/5 | 2/5 | 1 |

- **Step 3: conditional** = divide each row of the joint by its row marginal

| pr[ · \| row] | Sick | Not sick | Row sum |
|---|---|---|---|
| **Smoker** | 2/3 | 1/3 | 1 |
| **Non-smoker** | 1/2 | 1/2 | 1 |

- pr[Sick | Smoker] = 2/3, but pr[Sick | Non-smoker] = 1/2
- Code Sick = 1, not sick = 0: the **conditional expectation** E[Sick | Smoker] = 2/3 too

- **Step 4: independent?** If yes, every conditional row would equal the marginal pr[Sick] = 3/5
- 2/3 and 1/2 both differ from 3/5, so **not independent** in this sample
- Same check with the joint: pr[Smoker, Sick] = 2/5 = 0.40, but pr[Smoker] × pr[Sick] = 3/5 × 3/5 = 0.36
- **Direction matters**: pr[Smoker | Sick] divides the Sick **column** by 3/5: (2/5)/(3/5) = 2/3. It equals pr[Sick | Smoker] here only because both marginals are 3/5

# By Hand: Rain or Shine

Ten days of weather (S = sun, R = rain):

S S S R R R S S R R

- List the 9 (today, tomorrow) pairs and fill in the **joint** table of proportions
- Compute the **marginal** of today
- Compute the **conditional** PMF of tomorrow given today
- Is tomorrow **independent** of today?

| | | tomorrow | | |
|---|---|---|---|---|
| | | **S** | **R** | **Marginal** |
| today | **S** | ? | ? | ? |
| | **R** | ? | ? | ? |
| | **Marginal** | ? | ? | 1 |

**rows are the current observation** and **columns are the following observation**.

# With code: The Rain/Sun Cross Tab

```
df = pd.DataFrame({'today': weather})
df['tomorrow'] = df['today'].shift(-1)

# joint, with marginals
pd.crosstab(df['today'], df['tomorrow'], normalize=True,
margins=True)

# conditional
pd.crosstab(df['today'], df['tomorrow'], normalize='index')
```

- Shift the sequence by one to line up the 9 (today, tomorrow) pairs
- `normalize=True`: divide by all 9 pairs (**joint**). `margins=True` adds the **marginals**
- `normalize='index'`: divide each row by its row sum (**conditional**)

| today \ tomorrow | R | S | All |
|---|---|---|---|
| **R** | 0.3333 | 0.1111 | 0.4444 |
| **S** | 0.2222 | 0.3333 | 0.5556 |
| **All** | 0.5556 | 0.4444 | 1.0000 |

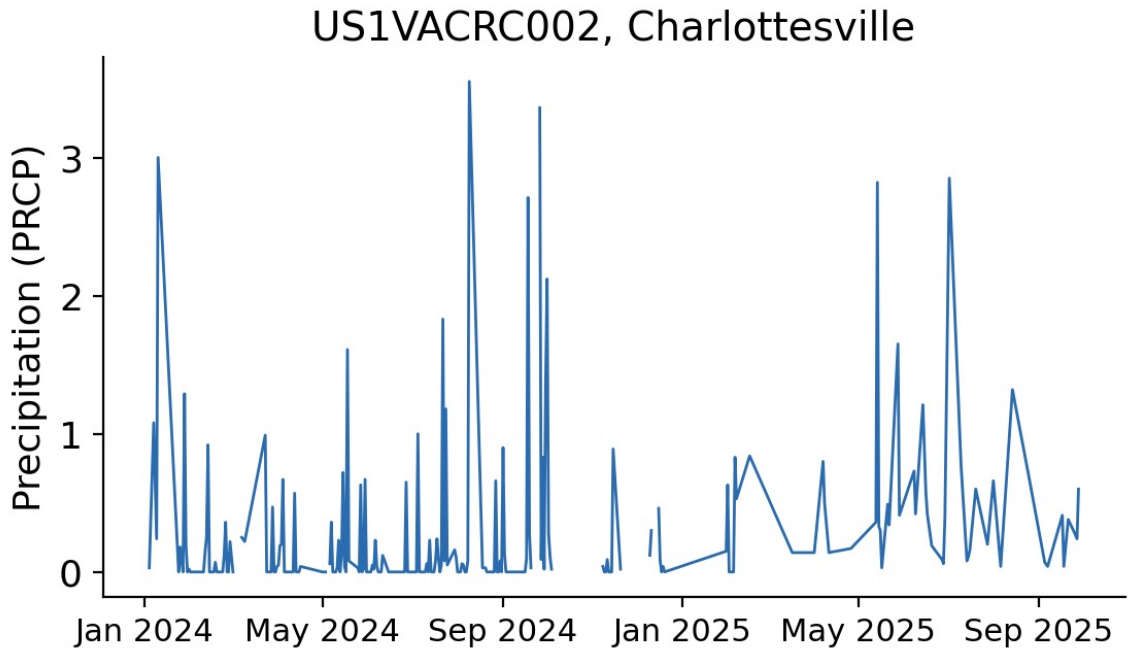
Joint, with margins

| today \ tomorrow | R | S |
|---|---|---|
| **R** | 0.7500 | 0.2500 |
| **S** | 0.4000 | 0.6000 |

Conditional: each **row** sums to 1. pandas sorts R before S

# Real Weather: Charlottesville, Dry or Rainy?

- One station, 271 recorded observations, Jan 2024 to Sep 2025
- **Dry** if precipitation $X_t = 0$, **Rainy** if $X_t > 0$
- t + 1 is the **next recorded observation**, not necessarily the next calendar day: the station's dates have gaps
- Missing precipitation (7 rows) stays missing, not Dry




```
df['y_t']   = df['PRCP_gtr0'].map({False:'Dry', True:'Rainy'})
df['y_tp1'] = df['PRCP_gtr0'].shift(-1).map({False:'Dry', True:'Rainy'})   # shift into the future
pd.crosstab(df['y_t'], df['y_tp1'])
```

`shift(-1)` places the following row's value beside the current row's value. Same trick as `weather[1:]` in the toy example

# Real Weather: Counts, Joint, and Marginals

| y_t \ y_t+1 | Dry | Rainy |
|---|---|---|
| **Dry** | 86 | 35 |
| **Rainy** | 37 | 98 |

Raw counts: 256 complete pairs

JOINT DISTRIBUTION

| y_t \ y_t+1 | Dry | Rainy | All |
|---|---|---|---|
| **Dry** | 0.336 | 0.137 | 0.473 |
| **Rainy** | 0.145 | 0.383 | 0.527 |
| **All** | 0.480 | 0.520 | 1.000 |

`normalize=True, margins=True`: the **whole table** sums to 1

CONDITIONAL PMF

| y_t \ y_t+1 | Dry | Rainy |
|---|---|---|
| **Dry** | 0.711 | 0.289 |
| **Rainy** | 0.274 | 0.726 |

`normalize='index'`: each **row** sums to 1

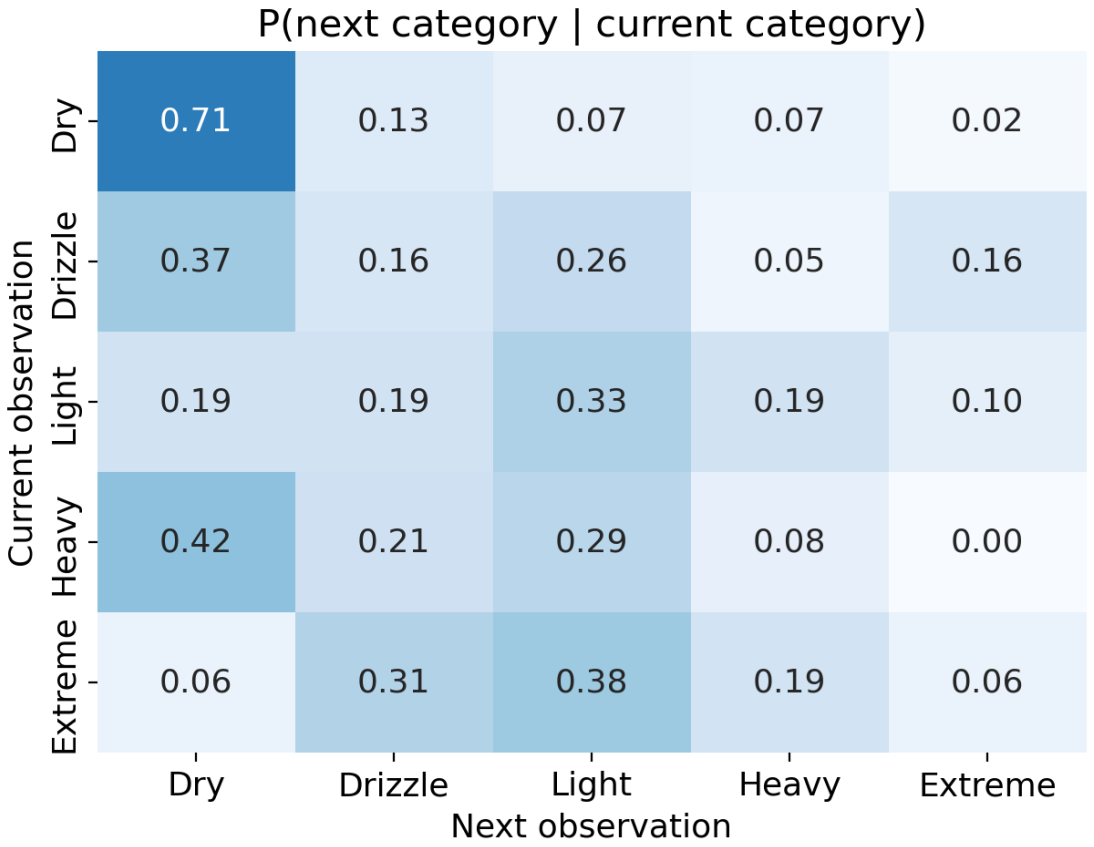
- Each cell in the joint PMF answers a question about **both** observations: 38% of pairs are rainy followed by rainy
- The conditional answers: **given** a rainy observation, 73% of the time the next one is rainy. Given dry, only 29%
- Is the next observation **independent** of the current one? Compare each conditional row with the marginal of $y_{t+1}$

# Real Weather: Five Categories

```
df['PRCP_bin'] = pd.cut(
    df['PRCP'],
    bins=[-0.001, 0, 0.1, 0.5, 1, np.inf],
    labels=['Dry', 'Drizzle', 'Light',
            'Heavy', 'Extreme'])
df['y_t']   = df['PRCP_bin']
df['y_tp1'] = df['PRCP_bin'].shift(-1)

pd.crosstab(df['y_t'], df['y_tp1'],
            normalize='index')
```

- **Binning** keeps some information about the amount: Dry/Rainy threw it away
- Read the Dry row: **given** a dry observation, the distribution of the next category
- Rare categories (16 Extreme, 24 Heavy pairs) give noisy rows




CONDITIONAL PMF

$$\widehat{P}(Y_{t+1} = j \mid Y_t = i) = \frac{\widehat{P}(Y_t = i, Y_{t+1} = j)}{\widehat{P}(Y_t = i)} = \frac{n_{ij}}{\sum_k n_{ik}}.$$

# Conditional Expectation: Tomorrow's Amount

- A conditional PMF gives probabilities of categories. A **conditional expectation** gives an **average numerical outcome** after restricting to a category:

$$E[X_{t+1} \mid Y_t = i]$$

```
df['PRCP_tp1'] = df['PRCP'].shift(-1)
df.groupby('y_t', observed=False)['PRCP_tp1'].describe()
```

- Average the **actual amounts** of the next observation, zeros included, not the bin labels
- A small mean can come with occasional large amounts: look at the count and spread too

| Current y_t | count | mean | std | max |
|---|---|---|---|---|
| **Dry** | 121 | 0.097 | 0.266 | 1.61 |
| **Drizzle** | 43 | 0.419 | 0.758 | 3.55 |
| **Light** | 52 | 0.448 | 0.705 | 3.00 |
| **Heavy** | 24 | 0.158 | 0.233 | 0.84 |
| **Extreme** | 16 | 0.389 | 0.519 | 2.12 |

$E[X_{t+1} \mid Y_t = i]$ is the **mean** column (precipitation units)

The **joint PMF** describes two categories together.

Summing its cells across a row/column gives a **marginal PMF**

Dividing a row by its marginal probability gives a **conditional PMF**.

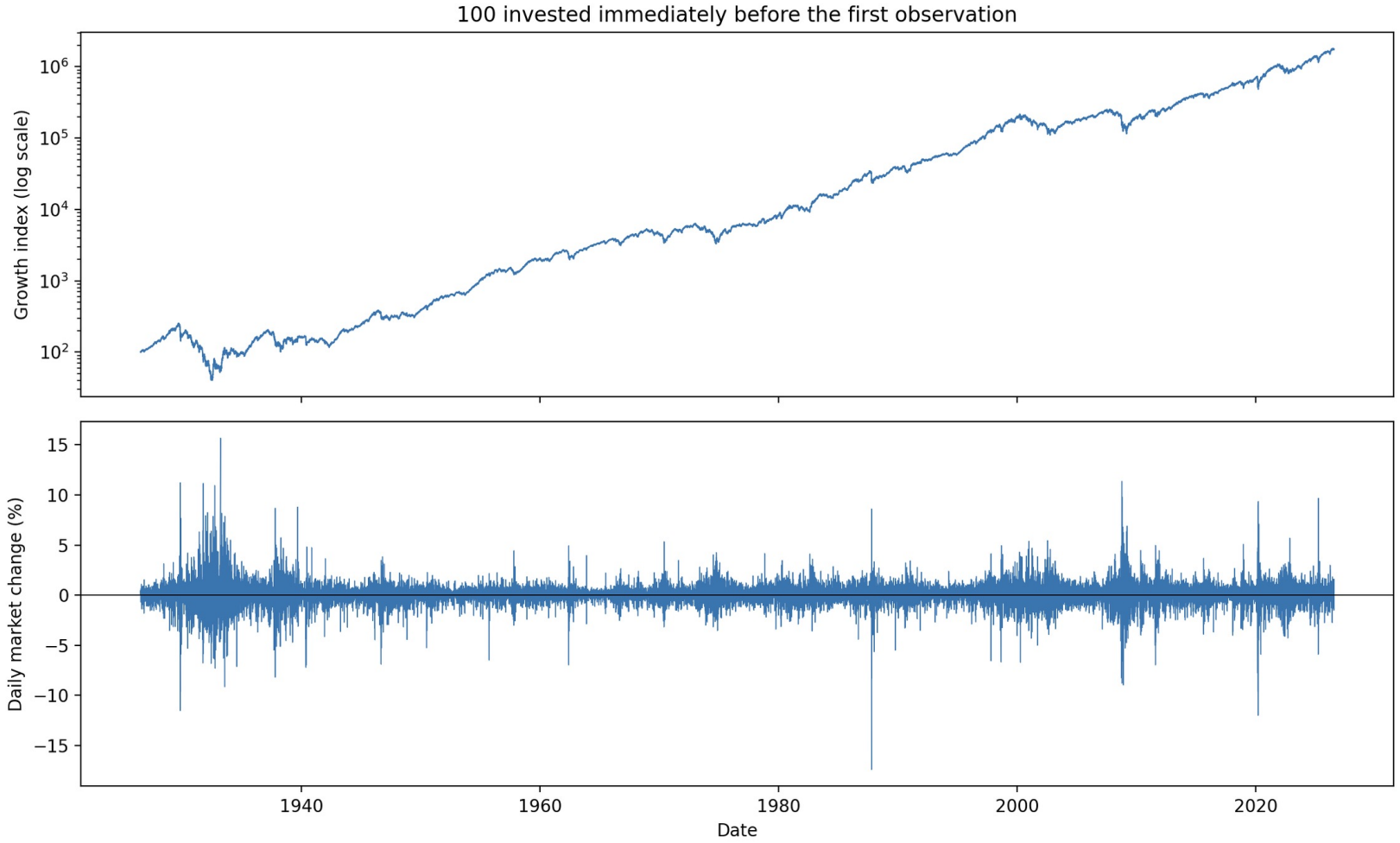
To obtain a **conditional expectation** of precipitation amount, we return to the **numerical** measurements and average within the selected current category.

For an indicator outcome, these ideas meet directly: the conditional expectation of "the following observation is rainy" equals the conditional probability that it is rainy. For precipitation amount, the conditional expectation is an average amount, including zeros.

# 2. Densities and Independence

# Market Returns: Up or Down Tomorrow?

- 26,296 daily US market returns, 1926 to 2026.
- Pair each trading day with the **next trading day**. Same two-category question as Dry/Wet




| direction_tp1 / direction_t | Down / flat | Up |
|---|---|---|
| **Down / flat** | 5909 | 5999 |
| **Up** | 5999 | 8388 |

# Market Returns: Up or Down Tomorrow?

- 26,296 daily US market returns, 1926 to 2026. Use the **daily percentage change**: $R_t$ = market_return_pct (2 means the market gained 2% that day)
- Pair each trading day with the **next trading day**. Same two-category question as Dry/Wet

| Today \ Next | Down / flat | Up | All |
|---|---|---|---|
| **Down / flat** | 0.225 | 0.228 | 0.453 |
| **Up** | 0.228 | 0.319 | 0.547 |
| **All** | 0.453 | 0.547 | 1.000 |

Joint with margins. The **whole table** sums to 1

| Today \ Next | Down / flat | Up |
|---|---|---|
| **Down / flat** | 0.496 | 0.504 |
| **Up** | 0.417 | 0.583 |

Conditional. Each **row** sums to 1

- Marginal: the market is up on 54.7% of days. Given today is up, tomorrow is up 58.3% of the time; given down, 50.4%
- The rows are close to the marginal but not equal. Is that a big difference?

# Independence

- Shrink the hexes to points and the counts become a **joint density**
- ($X_1$, ..., $X_n$) are **independent random variables** if the joint density decomposes into a product of marginals:

$$f_{X_1,X_2,...,X_n}(x_1, x_2, ..., x_n) = f_{X_1}(x_1) \times f_{X_2}(x_2) \times ... \times f_{X_n}(x_n)$$

- For tables, the same idea:

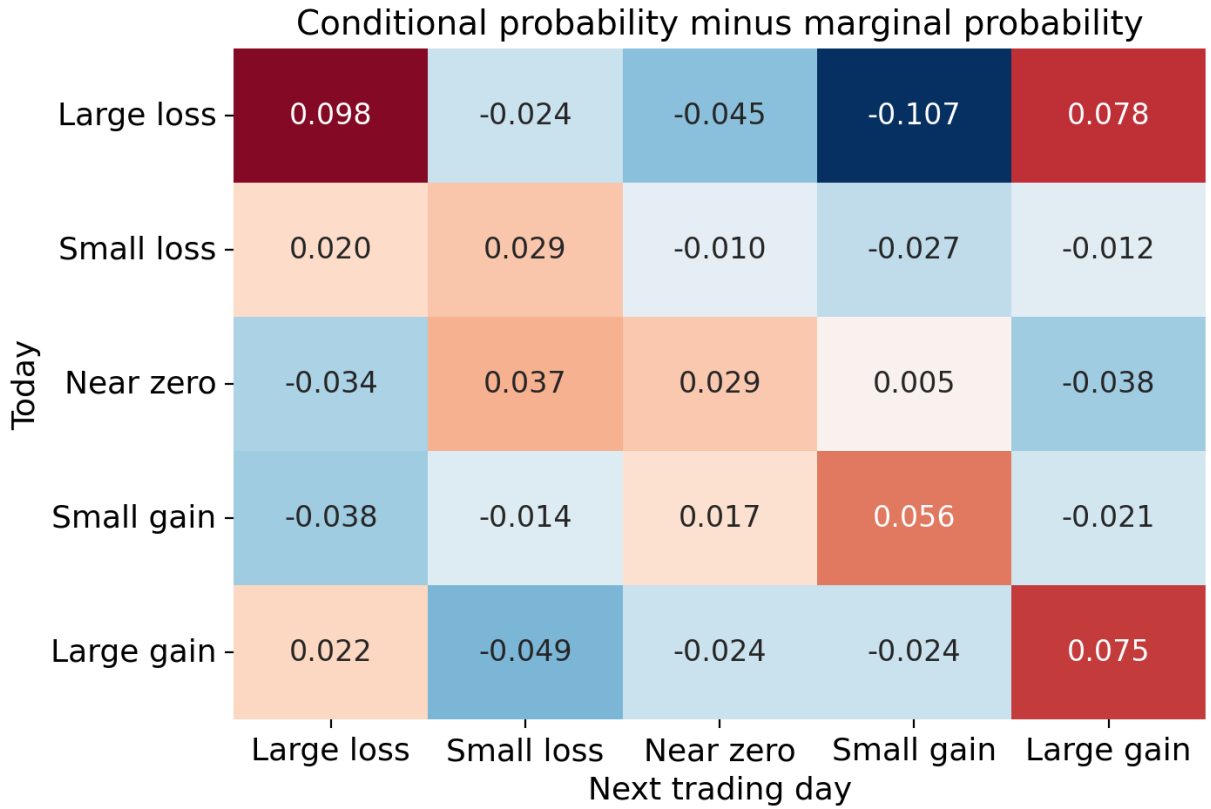
$$\text{pr}[Y = y | X = x] = \text{pr}[Y = y] \quad \text{for all } x, y$$

- If independent, **every conditional row equals the marginal**. Knowing X tells you nothing about Y

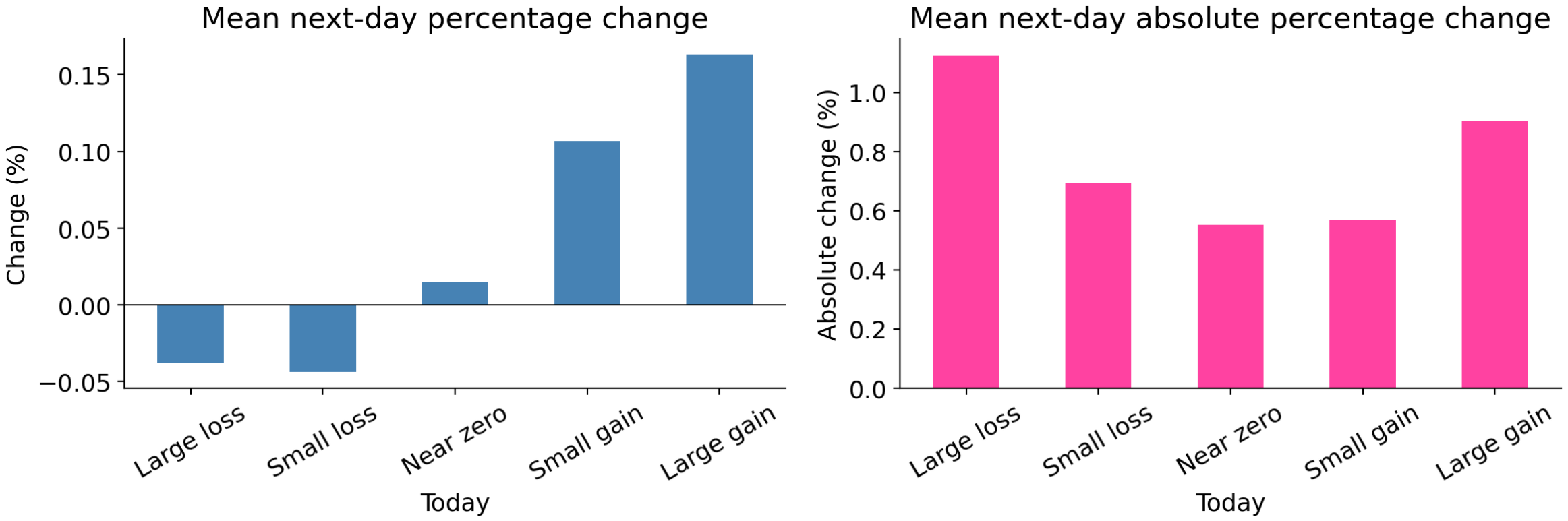
# Market Returns: Five Categories vs. Independence

- Bin daily percentage changes (%): **Large loss** ≤ –1 < **Small loss** ≤ –0.1 < **Near zero** ≤ 0.1 < **Small gain** ≤ 1 < **Large gain**
- Compute the conditional PMF of tomorrow's bin given today's, then **subtract the marginal** of tomorrow's bin
- If today and tomorrow were independent, every cell would be **0**
- Red: more likely than the marginal. Blue: less likely. Units are probability points (e.g. table is pr[tomorrow's bin | today's bin] – pr[tomorrow's bin] so a large loss tomorrow is **9.8 percentage points** more likely after a large loss today)

```
return_conditional.sub(next_category_pmf,
                       axis='columns')
```



# Conditional Expectation: Volatility Clustering

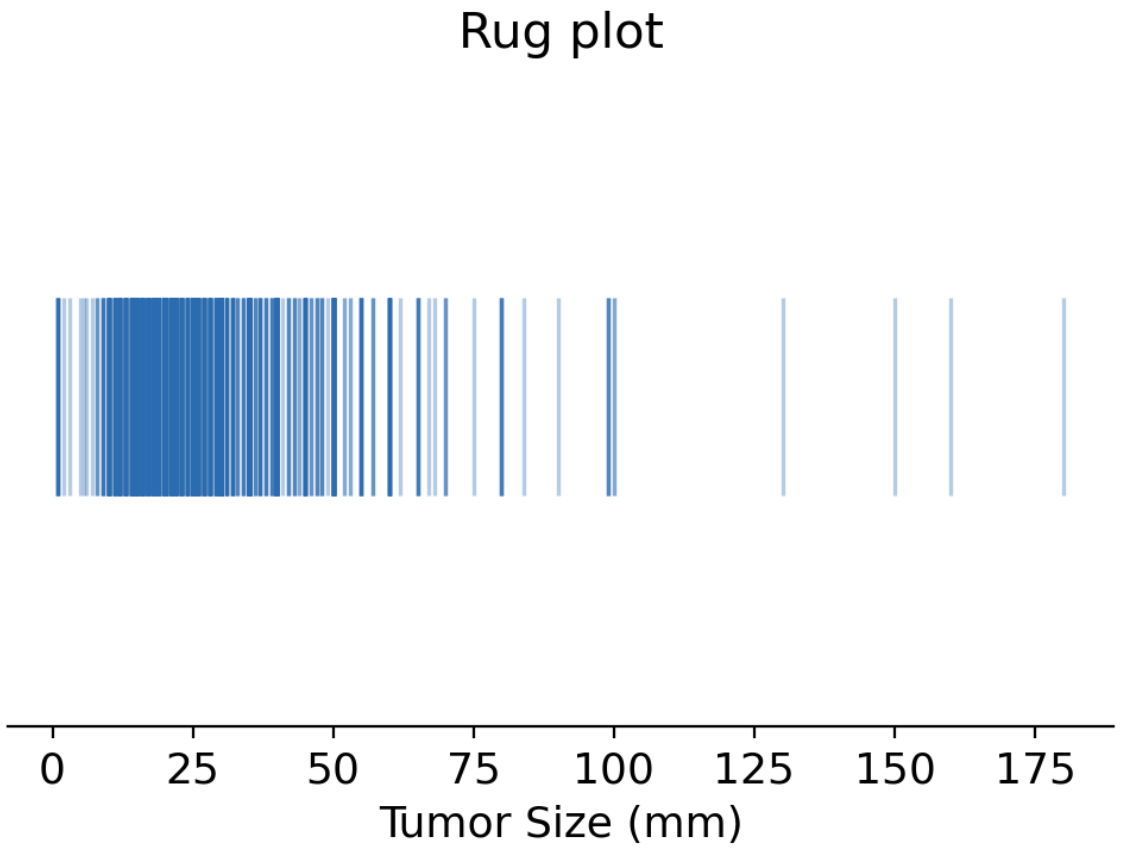


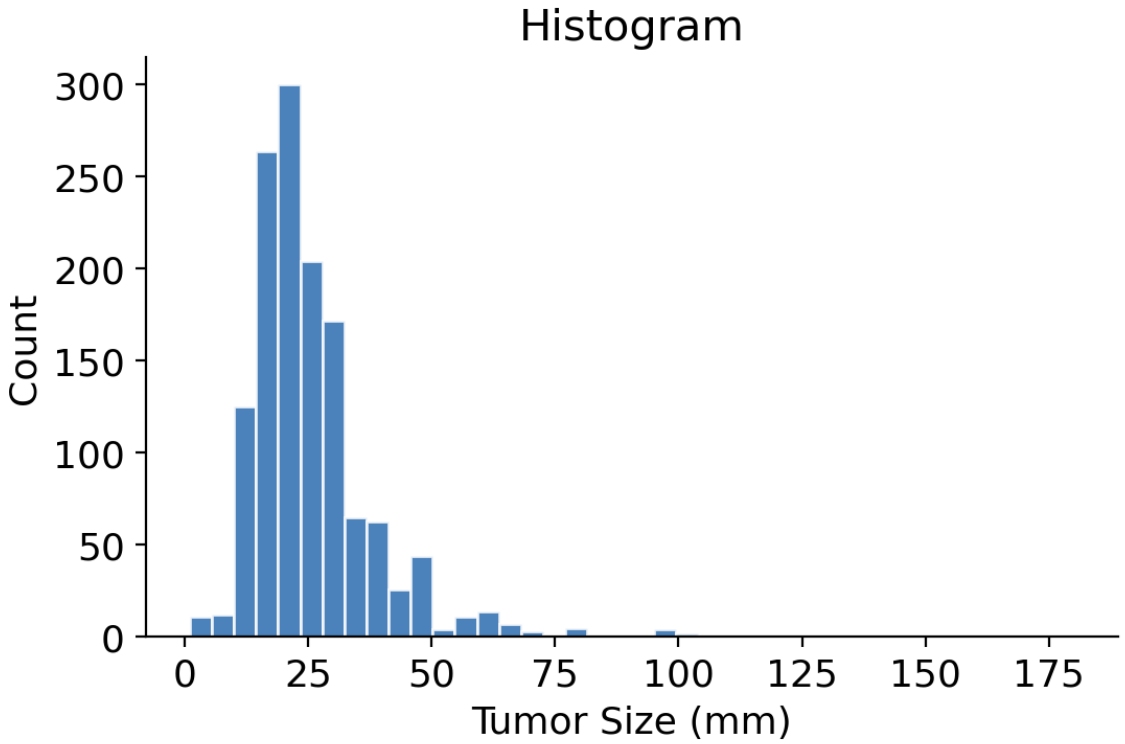

- $E[R_{t+1} \mid \text{today's bin}]$ (left) is close to 0 for every bin: direction is hard to predict
- $E[|R_{t+1}| \mid \text{today's bin}]$ (right): after a **large loss**, the next day's average move is 1.12%, about double the 0.55% after a near-zero day. Big moves follow big moves

# Points to Keep in Mind

- **Joint and conditional probabilities use different denominators.** The joint sums to 1 over the whole table. Each row of the conditional sums to 1 on its own
- **Conditional means retain the numerical outcome.** Category probabilities and average amounts answer different questions
- **Binning is a choice.** Different thresholds change the summaries. Rare categories give less precise estimates
- **These are descriptive summaries.** Differences between conditional rows describe association in this sample, not a causal effect

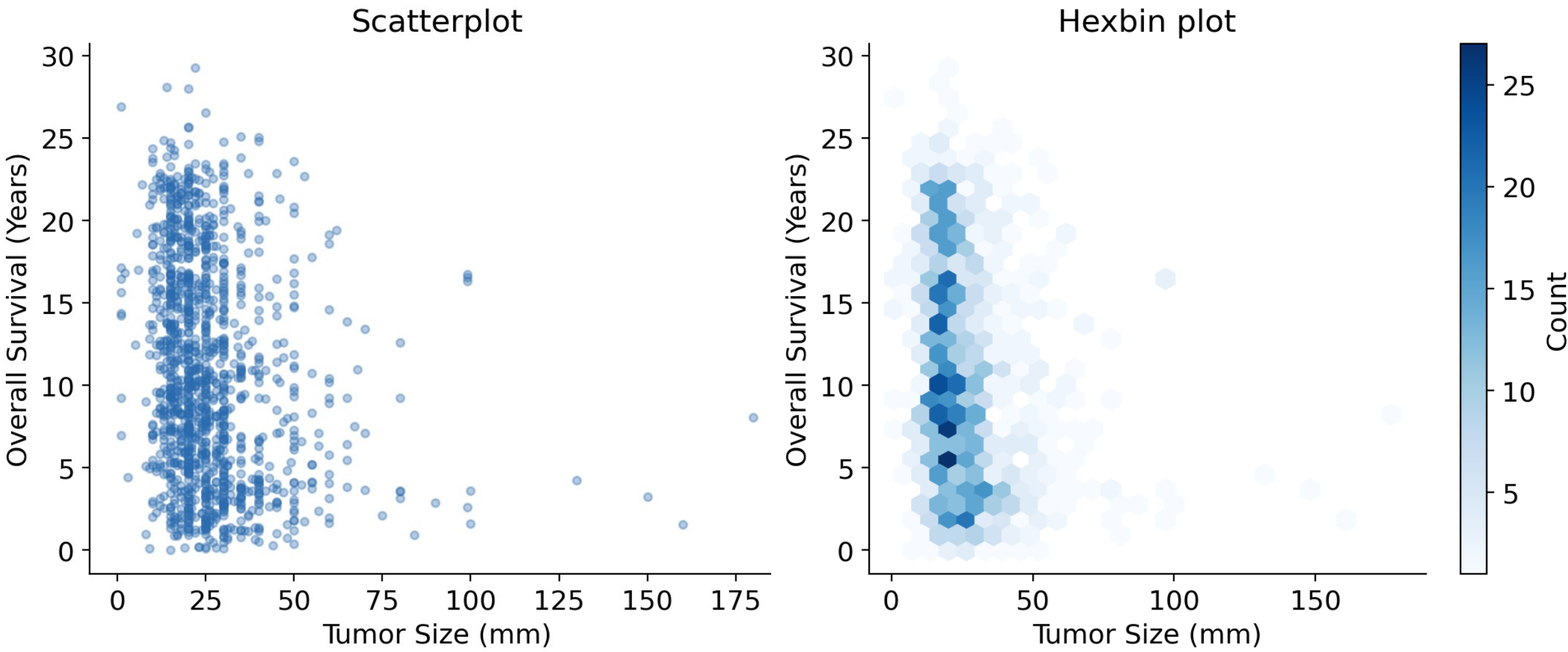
# Which Do You Like Better?




Same 1,343 tumor sizes. Where are most of the observations?

# Which Do You Like Better?




Tumor size vs. survival. A scatterplot is a two-dimensional rug plot: overlapping points hide where the mass is