# Notes for AP® Statistics Exam

Yuanqi Li
Humble Academy, Nanjing

## Contents

# 1 Describing Relationships

## 1.1 Scatterplots

A **scatterplot** displays the relationship between two quantitative variables measured on the same individuals. Mark values of one variable on the horizontal axis ($x$-axis) and values of the other variable on the vertical axis ($y$-axis). Plot each individual's data as a point on the graph.

**Definition 1.1: Scatterplot**

A **scatterplot** shows the relationship between two quantitative variables measured on the same individuals. The values of one variable appear on the horizontal axis, and the values of the other variable appear on the vertical axis. Each individual in the data appears as a point in the graph.

If we think that a variable $x$ may help explain, predict, or even cause changes in another variable $y$, we call $x$ an **explanatory variable** and $y$ a **response variable**. Always plot the explanatory variable, if there is one, on the $x$ axis of a scatterplot. Plot the response variable on the $y$ axis. If there is no explanatory-response distinction, either variable can go on the horizontal axis.

In examining a scatterplot, look for an overall pattern showing the **direction**, **form**, and **strength** of the relationship and then look for **outliers** or other departures from this pattern.

**Direction** If the relationship has a clear direction, we speak of either *positive association* or *negative association.*

**Form** *Linear relationships*, where the points show a straight-line pattern, are an important form of relationship between two variables. *Curved relationships* and clusters are other forms to watch for.

**Strength** The strength of a relationship is determined by *how close* the points in the scatterplot lie to a simple form such as a line.

## 1.2 Correlation

The **correlation** $r \in [-1, +1]$ measures the *strength* and *direction* of the *linear association* between two quantitative variables $x$ and $y$. Although you can calculate a correlation for *any* scatterplot, $r$ measures strength for *only* straight-line relationships.

Correlation indicates the direction of a linear relationship by its sign and magnitude:

- $r > 0$ for a positive association;
- $r < 0$ for a negative association;
- $r = \pm 1$ occurs only when the points on a scatterplot lie exactly on a straight line; and
- $|r|$ indicates the strength of a linear relationship by how close it is to -1 or +1.

## Definition 1.2: Correlation

The **correlation** $r$ measures the direction and strength of the linear relationship between two quantitative variables.
Suppose that we have data on variables $x$ and $y$ for $n$ individuals given as:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n).$$

The means and standard deviations of the two variables are $\bar{x}$ and $s_x$ for the $x$-values, and $\bar{y}$ and $s_y$ for the $y$-values. The correlation $r$ between $x$ and $y$ is given by

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = \frac{1}{n-1} \sum_{i=1}^{n} z_{x_i} z_{y_i}.$$

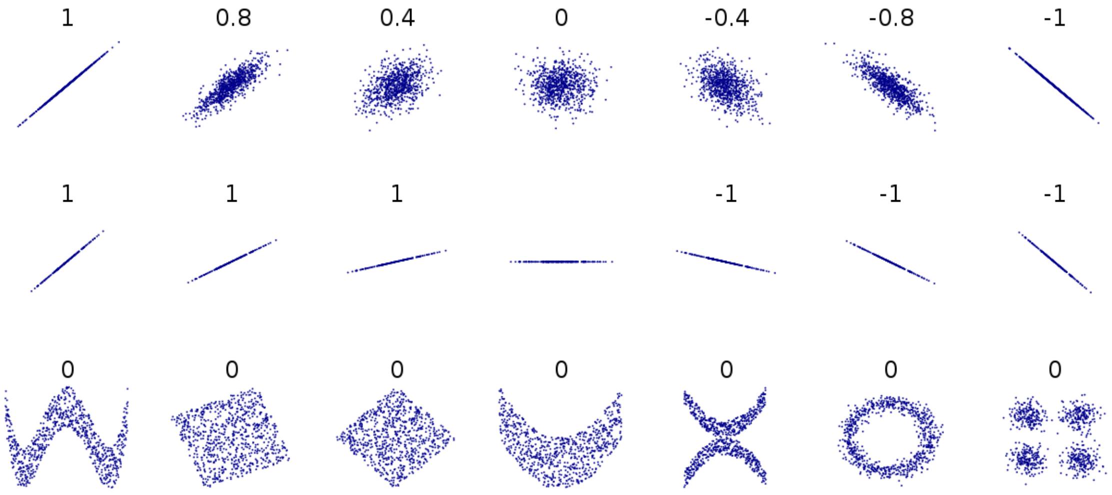

**Figure 1:** Several sets of $(x, y)$ points, with the correlation coefficient of $x$ and $y$ for each set. The correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of $y$'s is zero.

How correlation behaves is more important than the details of the formula. Here's what you need to know in order to interpret correlation correctly.

- Correlation makes no distinction between explanatory and response variables.
- Because $r$ uses the standardized values of the observations, $r$ does not change when we change the units of measurement of $x$, $y$, or both.
- The correlation $r$ itself has no unit of measurement.

Describing the relationship between two variables is more complex than describing the distribution of one variable. Here are some cautions to keep in mind when you use correlation.

- Correlation does not imply causation.
- Correlation requires that both variables be quantitative, so that it makes sense to do the arithmetic indicated by the formula for $r$.

- Correlation only measures the strength of a *linear relationship* between two variables, i.e., it does not describe *curved relationships* between variables, no matter how strong the relationship is.
- A value of $r$ close to 1 or $-1$ does not guarantee a linear relationship between two variables, e.g., when the underlying relationship is curved but not linear.
- Like mean and standard deviation, the correlation is not resistant: $r$ is strongly affected by a few outlying observations. Use $r$ with caution when outliers appear in the scatterplot.
- Correlation is not a complete summary of two-variable data, even when the relationship between the variables is linear. You should give the means and standard deviations of both $x$ and $y$ along with the correlation.

## 1.3 Linear Regression

A **regression** line is a *model* for the data. It summarizes the relationship between two variables, but only in a specific setting: when one of the variables helps explain or predict the other. Regression, unlike correlation, requires that we have an explanatory variable and a response variable.

**Definition 1.3: Regression line**

A **regression line** is a line that describes how a response variable $y$ changes as an explanatory variable $x$ changes. We often use a regression line to predict the value of $y$ for a given value of $x$.
Suppose that $y$ is a response variable (plotted on the vertical axis) and $x$ is an explanatory variable (plotted on the horizontal axis). A regression line relating $y$ to $x$ has an equation of the form

$$\hat{y} = a + bx,$$

where

- $\hat{y}$ (read "$y$ hat") is the **predicted value** of the response variable $y$ for a given value of the explanatory variable $x$;
- $b$ is the **slope**, the amount by which $y$ is predicted to change when $x$ increases by one unit; and
- $a$ is the **y-intercept** the predicted value of $y$ when $x = 0$.

Few relationships are linear for *all* values of the explanatory variable. Don't make predictions using values of $x$ that are *much larger* or *much smaller* than those that actually appear in your data.

A good regression line makes the vertical deviations of the points from the line as small as possible. These vertical deviations represent "leftover" variation in the response variable after fitting the regression line. For that reason, they are called **residuals**.

### Definition 1.4: Residual

A **residual** is the difference between an observed value of the response variable and the value predicted by the regression line, i.e.,

$$\begin{aligned} \text{residual} &= \text{observed } y - \text{predicted } y \\ &= y - \hat{y}. \end{aligned}$$

The regression line we want is the one that minimizes the sum of the squared residuals.

### Definition 1.5: Least-squares regression line

The **least-squares regression line** of $y$ on $x$ is the line that makes the sum of the squared residuals as small as possible.
Suppose that we have data on variables $x$ and $y$ for $n$ individuals given as:

$$(x_1, y_1), (x_2, y_2), ..., (x_n, y_n).$$

The means and standard deviations of the two variables are $\bar{x}$ and $s_x$ for the $x$-values, and $\bar{y}$ and $s_y$ for the $y$-values. The least-squares regression line is the line $\hat{y} = a + bx$ where

$$b = r\frac{s_y}{s_x} \quad \text{and} \quad a = \bar{y} - b\bar{x}.$$

Note that when displaying the equation of a least-squares regression line, the calculator will report the slope and intercept with much more precision than we need. However, there is no firm rule for how many decimal places to show for answers on the AP exam. The advice is that decide how much to round based on the context of the problem you are working on.

Although residuals can be calculated from any model that is fitted to the data, the residuals from the least-squares line have a special property: the mean of the least-squares residuals is *always* zero.

Correlation and regression are powerful tools for describing the relationship between two variables. When you use these tools, you should be aware of their limitations.

- The distinction between explanatory and response variables is important in regression.
- Correlation and regression lines describe only linear relationships, i.e., the results are useful only if the scatterplot shows a linear pattern. Always plot your data. See also Figure 2.
- Correlation and least-squares regression lines are not resistant. One unusual point in a scatterplot can greatly change the value of $r$. Least-squares lines make the sum of the squares of the vertical distances to the points as small as possible. A point that is extreme in the $x$ direction with no other points near it pulls the line toward itself. We call such points **influential**. See also Figure 2 and 3.
- Association does not imply causation. A strong association between two variables is not enough to draw conclusions about cause and effect.

### Definition 1.6: Outliers and influential observations in regression

An **outlier** is an observation that lies outside the overall pattern of the other observations. Points that are outliers in the $y$ direction but not the $x$ direction of a scatterplot have large residuals. Other outliers may not have large residuals.
An observation is **influential** for a statistical calculation if removing it would markedly change the result of the calculation. Points that are outliers in the $x$ direction of a scatterplot are often influential for the least-squares regression line.

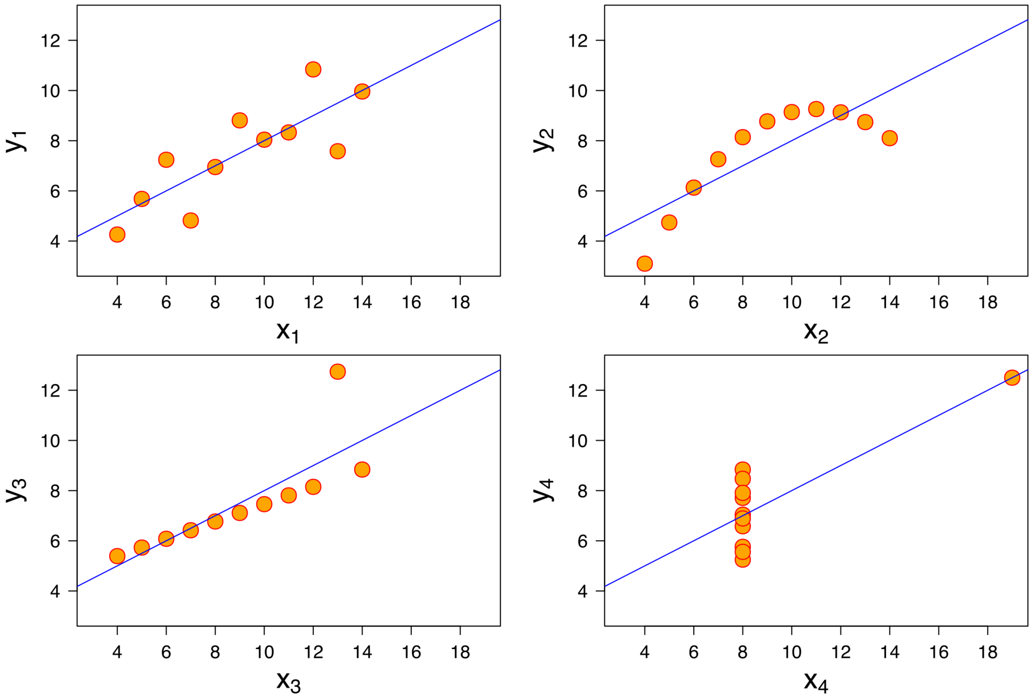

**Figure 2:** All four sets are identical when examined using simple summary statistics, but vary considerably when graphed. $\bar{x} = 9$, $\bar{y} = 7.50$, $s_x^2 = 11$, $s_y^2 = 4.125$, correlation $r = 0.816$, linear regression line $\hat{y} = 3.00 + 0.500x$, coefficient of determination $r^2 = 0.67$.

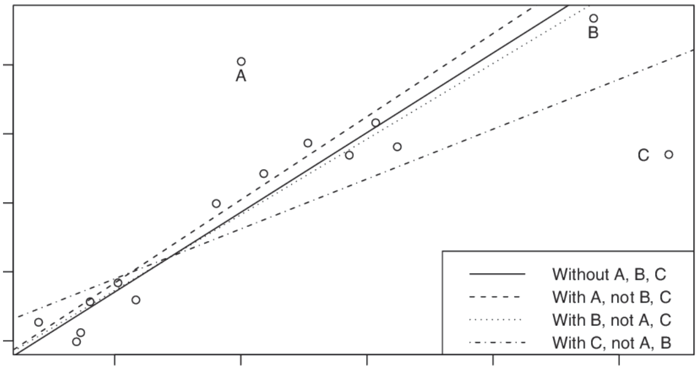

**Figure 3:** Outliers in regression are observations that fall far from the "cloud" of points. These points are especially important because they can have a strong influence on the least squares line.

### 1.3.1 Residual Plot

A **residual plot** in effect turns the regression line horizontal. It magnifies the deviations of the points from the line, making it easier to see unusual observations and patterns.

**Definition 1.7: Residual plot**

A **residual plot** is a scatterplot of the residuals against the explanatory variable. Residual plots help us assess whether a linear model is appropriate.

When an obvious curved pattern exists in a residual plot, the model we are using is not appropriate. Remember when we calculate a residual, we are calculating what is left over after subtracting the predicted value from the observed value:

$$\text{residual} = \text{observed } y - \text{predicted } y.$$

Likewise, when we look at the form of a residual plot, we are looking at the form that is left over after subtracting the form of the model from the form of the association:

$$\text{form of residual plot} = \text{form of association} - \text{form of model}.$$

When there is a leftover form in the residual plot, the form of the association and form of the model are not the same. However, if the form of the association and form of the model are the same, the residual plot should have no form, other than *random scatter*.

### 1.3.2 The Standard Deviation of the Residuals: $s$

To assess how well the line fits all the data, we need to consider the residuals for each of the predictions we made, not just one. Using these residuals, we can estimate the "typical" prediction error when using the least-squares regression line. To do this, we calculate the standard deviation of the residuals.

**Definition 1.8: Standard deviation of the residuals**

If we use a least-squares line to predict the values of a response variable $y$ from an explanatory variable $x$, the **standard deviation of the residuals** ($s$) is given by

$$s = \sqrt{\frac{\sum \text{residuals}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y})^2}{n-2}}.$$

This value gives the approximate size of a *typical prediction error*, i.e., residual.

### 1.3.3 The Coefficient of Determination: $r^2$

If we don't know the value for the additional explanatory variable, we can't use the regression line to make a prediction. What should we do? Our best strategy is to use the mean value of the response variable of the other points as our prediction. However, if we learn the value of the new explanatory variable $x$, then we could use the least-squares line to make prediction. How much better does the regression line do at prediction than simply using the average $\bar{y}$?

### Definition 1.9: The coefficient of determination

The **coefficient of determination** $r^2$ is the fraction of the variation in the values of $y$ that is accounted for by the least-squares regression line of $y$ on $x$. We can calculate $r^2$ using the following formula:

$$r^2 = 1 - \frac{\sum \text{residuals}^2}{\sum (y_i - \bar{y})^2}.$$

If all the points fall directly on the least-squares line, the sum of squared residuals is 0 and $r^2 = 1$. Then all the variation in $y$ is accounted for by the linear relationship with $x$. Because the least-squares line yields the smallest possible sum of squared prediction errors, the sum of squared residuals can never be more than the sum of squared deviations from the mean of $y$. In the worst-case scenario, the least-squares line does no better at predicting $y$ than $y = \bar{y}$ does. Then the two sums of squares are the same and $r^2 = 0$.

It seems fairly remarkable that the coefficient of determination is actually the correlation squared. This fact provides an important connection between correlation and regression. When you see a correlation, square it to get a better feel for how well the least-squares line fits the data.

# 2 Probability Fundamentals

**Chance process** A chance process has outcomes that we cannot predict but that nonetheless have a regular distribution in very many repetitions. **The law of large numbers** says that the proportion of times that a particular outcome occurs in many repetitions will approach a single number. This long-run relative frequency of a chance outcome is its **probability**. A probability is a number between 0 (never occurs) and 1 (always occurs).

Probabilities describe only what happens in the long run. Short runs of random phenomena like tossing coins or shooting a basketball often don't look random to us because they do not show the regularity that emerges only in very many repetitions.

**Simulation** A simulation is an imitation of chance behavior, most often carried out with random numbers. To perform a simulation, follow the four-step process:

**State** Ask a question of interest about some chance process.

**Plan** Describe how to use a chance device to imitate one repetition of the process. Tell what you will record at the end of each repetition.

**Do** Perform many repetitions of the simulation.

**Conclude** Use the results of your simulation to answer the question of interest.

## 2.1 Basic Probability Rules

**Sample space** A **probability model** describes chance behavior by listing the possible outcomes in the **sample space** $S$ and giving the probability that each outcome occurs.

**Events** An **event** is a subset of the possible outcomes in the sample space. To find the probability that an event $A$ happens, $\mathbf{Pr}(A)$, we can rely on some basic probability rules:

- For any event $A$, $0 \leq \mathbf{Pr}(A) \leq 1$.
- $\mathbf{Pr}(S) = 1$, where $S$ is the sample space.
- If all outcomes in the sample space are equally likely, then

$$\mathbf{Pr}(A) = \frac{\text{\# outcomes corresponding to event } A}{\text{\# outcomes in the sample space}}.$$

## 2.2 Probability Involving Multiple Events

A **two-way table** or a **Venn diagram** can be used to display the sample space for a chance process. Two-way tables and Venn diagrams can also be used to find probabilities involving events $A$ and $B$. For example, event $A \cup B$ ($A$ or $B$) consists of all outcomes in event $A$, event $B$, or both. Event $A \cap B$ ($A$ and $B$) consists of all outcomes in both $A$ and $B$.

**Complement rule** Event $A^C$ ($A$ not happen) consists of all outcomes not in event $A$, whose probability is given by

$$\mathbf{Pr}\left(A^C\right) = 1 - \mathbf{Pr}(A).$$

**Addition rule** The general addition rule states that the probability of either A or B, or both, happens is

$$\mathbf{Pr}(A \text{ or } B) = \mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B) - \mathbf{Pr}(A \cap B).$$

In the special case that $A$ and $B$ are *mutually exclusive* (disjoint), i.e., they have no outcomes in common, the addition rule becomes

$$\mathbf{Pr}(A \text{ or } B) = \mathbf{Pr}(A \cup B) = \mathbf{Pr}(A) + \mathbf{Pr}(B).$$

## 2.3 Conditional Probability

**Conditional probability** If one event $B$ has happened, the chance that another event $A$ will happen is a **conditional probability**, denoted $\mathbf{Pr}(A \mid B)$, which can be calculated by

$$\mathbf{Pr}(A \mid B) = \frac{\mathbf{Pr}(A \cap B)}{\mathbf{Pr}(B)}.$$

**Multiplication rule** The general multiplication rule states that the probability of events $A$ and $B$ occurring together is

$$\mathbf{Pr}(A \text{ and } B) = \mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\,\mathbf{Pr}(B \mid A).$$

In the special case of *independent* events, the multiplication rule becomes

$$\mathbf{Pr}(A \text{ and } B) = \mathbf{Pr}(A \cap B) = \mathbf{Pr}(A)\,\mathbf{Pr}(B).$$

**The law of total probability** Let events $B_1, B_2, ..., B_n$ be a **partition** of the sample space, i.e., they are disjoint and their union is the entire sample space. Then

$$\mathbf{Pr}(A) = \sum_{i=1}^{n} \mathbf{Pr}(A \mid B_i)\,\mathbf{Pr}(B_i).$$

**Bayes' theorem** Bayes' theorem gives a way of *inverting* conditions:

$$\mathbf{Pr}(A \mid B) = \frac{\mathbf{Pr}(B \mid A)\,\mathbf{Pr}(A)}{\mathbf{Pr}(B)},$$

where

- $\mathbf{Pr}(A \mid B)$ is the probability of event $A$ occurring given $B$ has already happened, a.k.a., the **posterior** of $A$ given $B$.
- $\mathbf{Pr}(B \mid A)$ is the probability of event $B$ occurring given $A$ has already happened.
- $\mathbf{Pr}(A)$ and $\mathbf{Pr}(B)$ are the probabilities of observing $A$ and $B$ respectively without given any conditions, a.k.a., **marginal probability** or **prior probability**. They can often be found using the law of total probability.

## 2.4 Mutually Exclusive Events vs. Independent Events

**Mutually exclusive events** Events are **mutually exclusive** if the occurrence of one event excludes the occurrence of the other(s). Mutually exclusive events cannot happen at the same time. For example, when tossing a coin, the result can either be heads or tails but cannot be both.

**Independent events** Events are **independent** if the occurrence of one event does not influence (and is not influenced by) the occurrence of the other(s). For example, when tossing two coins, the result of one flip does not affect the result of the other. Formally, two events are independent iff one of the following equivalent statements holds:

$$\begin{aligned} \mathbf{Pr}(A \cap B) &= \mathbf{Pr}(A)\,\mathbf{Pr}(B) \\ \mathbf{Pr}(A \mid B) &= \mathbf{Pr}(A) \\ \mathbf{Pr}(B \mid A) &= \mathbf{Pr}(B) \end{aligned}$$

In a word, *mutually exclusive events are not independent, and independent events are not mutually exclusive.*

| | Mutually exclusive | Independent |
|---|---|---|
| $\mathbf{Pr}(A \cap B)$ | $0$ | $\mathbf{Pr}(A)\,\mathbf{Pr}(B)$ |
| $\mathbf{Pr}(A \cup B)$ | $\mathbf{Pr}(A) + \mathbf{Pr}(B)$ | $\mathbf{Pr}(A) + \mathbf{Pr}(B) - \mathbf{Pr}(A)\,\mathbf{Pr}(B)$ |
| $\mathbf{Pr}(A \mid B)$ | $0$ | $\mathbf{Pr}(A)$ |
| $\mathbf{Pr}\left(A \mid B^C\right)$ | $\dfrac{\mathbf{Pr}(A)}{1 - \mathbf{Pr}(B)}$ | $\mathbf{Pr}(A)$ |

# 3 Random Variables

**Random variable** A random variable takes numerical values determined by the outcome of a chance process. The **probability distribution** of a random variable $X$ tells us what the possible values of $X$ are and how probabilities are assigned to those values. There are two types of random variables: discrete and continuous.

**Discrete** A discrete random variable has a *fixed set of possible values* with gaps between them. The probability distribution assigns each of these values a probability between 0 and 1 such that the sum of all the probabilities is exactly 1. The probability of any event is the sum of the probabilities of all the values that make up the event.

**Continuous** A continuous random variable takes all values in some *interval of numbers.* A density curve describes the probability distribution of a continuous random variable. The probability of any event is the area under the curve above the values that make up the event.

**Expected value** Expected value (a.k.a., *mean*, *expectation*, or *average*) is a weighted average of the possible outcomes of a random variable. Mathematically, if $x_1, x_2, ..., x_n$ are all of the distinct possible values that a discrete random variable $X$ can take, the expected value of $X$ is given by

$$\mu_X = \mathbf{E}[X] = \sum_{i=1}^{n} x_i \, \mathbf{Pr}(X = x_i).$$

**Linearity of expected value** For any random variables $X$ and $Y$, and constants $a$, $b$, and $c$,

$$\mathbf{E}[aX + bY + c] = a\,\mathbf{E}[X] + b\,\mathbf{E}[Y] + c.$$

**Variance and standard deviation** For a random variable $X$, its variance $\delta_X^2$ is the *average squared deviation* of the values of the variable from their mean; its standard deviation $\delta_X$ is the square root of the variance, measuring the typical distance of the values in the distribution from the mean. For a discrete random variable $X$,

$$\begin{aligned}\delta_X^2 &= \mathrm{Var}[X] = \sum_{i=1}^{n} (x_i - \mu_X)^2 \, \mathbf{Pr}(X = x_i) \\ \delta_X &= \sqrt{\delta_X^2}\end{aligned}$$

## 3.1 Transforming Random Variables

**Addition and subtraction** Adding a positive constant $a$ to (or subtracting $a$ from) a random variable increases (decreases) the mean of the random variable by $a$ but does not affect its standard deviation or the shape of its probability distribution.

**Multiplication and division** Multiplying (dividing) a random variable by a positive constant $b$ multiplies (divides) the mean of the random variable by $b$ and the standard deviation by $b$ but does not change the shape of its probability distribution.

|  | Center and location<br>*Mean, median, quartiles, percentiles* | Spread<br>*range, IQR, standard deviation* | Shape |
|---|---|---|---|
| $X+a$ | increased by $a$ | No change | No change |
| $X-a$ | decreased by $a$ | No change | No change |
| $bX$ | multiplied by $b$ | multiplied by $b$ | No change |
| $X/b$ | divided by $b$ | divided by $b$ | No change |

**Linear transformation** A linear transformation of a random variable involves adding or subtracting a constant $a$, multiplying or dividing by a constant $b$, or both. We can write a linear transformation of the random variable $X$ in the form $Y = a + bX$. The shape, center, and spread of the probability distribution of $Y$ are as follows:

**Shape** Same as the probability distribution of $X$ if $b > 0$

**Center** $\mu_Y = a + b\mu_X$

**Spread** $\delta_Y = |b|\delta_X$

### 3.2 Combining Random Variables

**Mean of combined random variables** If $X$ and $Y$ are *any* two random variables,

$$\mu_{X\pm Y} = \mu_X \pm \mu_Y.$$

**Variance of combined random variables** If $X$ and $Y$ are two *independent* random variables,

$$\delta^2_{X\pm Y} = \mu_X + \mu_Y.$$

Note that the variance of either the sum or the difference of two independent variables is *always* the sum of their variances.

**Combined normal random variables** Let $X$ and $Y$ be independent random variables that are normally distributed, then their sum is also normally distributed, i.e., if

$$X \sim N\left(\mu_X, \delta^2_X\right), \quad Y \sim N\left(\mu_Y, \delta^2_Y\right), \quad Z = X + Y,$$

then

$$Z \sim N\left(\mu_X + \mu_Y, \delta^2_X + \delta^2_Y\right)$$

### 3.3 Binomial Distribution

**Binomial coefficient** The binomial coefficient counts the number of ways $k$ successes can be arranged among $n$ trials, calculated by

$$\binom{n}{k} = \frac{n!}{k!(n-k)!},$$

where the factorial of $n$ is

$$n! = n(n-1)(n-2)\cdots(3)(2)(1),$$

for positive whole numbers $n$ and $0! = 1$.

**Binomial settings** A binomial setting consists of $n$ independent trails of the same chance process, each resulting in a success or a failure, with probability of success $p$ on each trial. The count $X$ of successes is a **binomial random variable**. Its probability distribution is a **binomial distribution**.

**Binomial distribution** If $X$ has the binomial distribution with parameters $n$ and $p$, i.e., $X \sim \text{Binomial}(n, p)$, the possible values of $X$ are the whole numbers $0, 1, ..., n$. Its probability mass function (pmf), mean, and variance are given by

**pmf** $\mathbf{Pr}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$

**Mean** $\mu_X = np$

**Variance** $\delta_X^2 = np(1-p)$

**Shape of binomial distribution** The shape of the binomial distribution is directly controlled by $n$ – the number of independent trials, and $p$ – the probability of success.

**Small p, small n** The binomial distribution is *skewed right*, i.e., the bulk of the probability falls in the smaller numbers $0, 1, 2, ...$, and the distribution tails off to the right. See Figure 4 ($n = 15, p = 0.1$) and ($n = 15, p = 0.2$).

**Large p, small n** The binomial distribution is *skewed left*, i.e., the bulk of the probability falls in the smaller numbers $n, n-1, n-2, ...$, and the distribution tails off to the left. See Figure 4 ($n = 15, p = 0.8$) and ($n = 15, p = 0.9$).

**p = 0.5, all n** The binomial distribution is *symmetric.* See Figure 4 ($n = 15, p = 0.5$) and ($n = 40, p = 0.5$).

**All p, large n** The binomial distribution *approaches symmetry.* For example, if $p = 0.2$ and $n$ is small, we'd expect the binomial distribution to be skewed to the right. For large $n$, however, the distribution is *nearly symmetric.* See Figure 4 ($n = 40, p = 0.1$), ($n = 40, p = 0.2$), ($n = 40, p = 0.8$), ($n = 40, p = 0.9$).

### 3.4 Geometric Distribution

**Geometric settings** A geometric setting consists of repeated trials of the same chance process in which the probability $p$ of success is the same on each trial, and the goal is to count the number of trials it takes to get one success. If $Y$ is the number of trials required to obtain the first success, then $Y$ is a geometric random variable. Its probability distribution is called a geometric distribution.

**Geometric distribution** If $Y$ has the geometric distribution with probability of success $p$, i.e., $Y \sim \text{Geometric}(p)$, the possible values of $Y$ are the positive integers $1, 2, 3, ....$ Its probability mass function (pmf), mean, and variance are given by

**pmf** $\mathbf{Pr}(Y = k) = (1-p)^{k-1} p$

**Mean** $\mu_Y = 1/p$

**Variance** $\delta_Y^2 = (1-p)/p^2$

**Shape of binomial distribution** The shape of the geometric distribution is directly controlled by $p$ – the probability of success. The larger the $p$, the steeper the function in the beginning. Note that the function is *monotonically decreasing.*

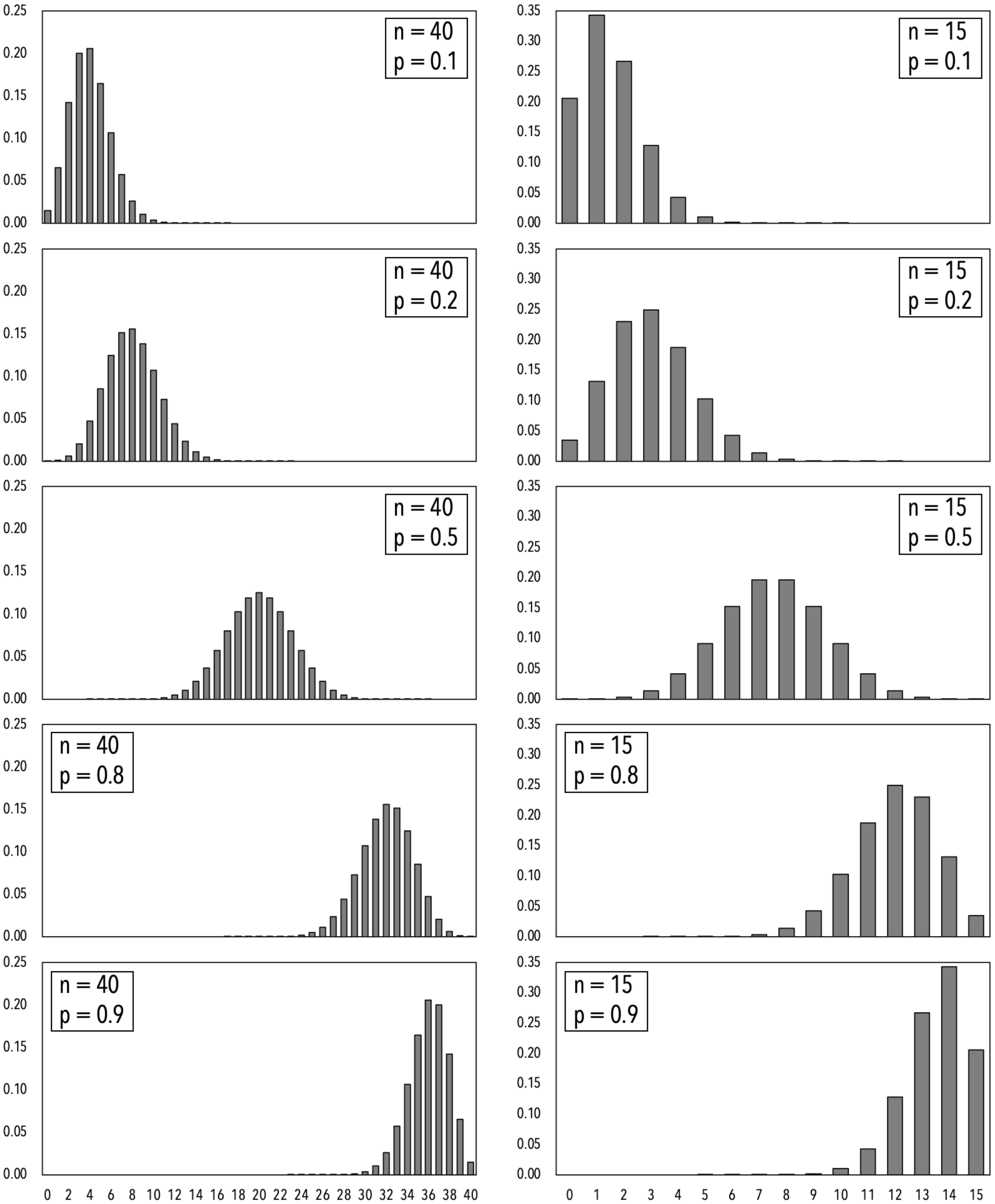

**Figure 4:** Comparison of binomial distribution of different parameters.

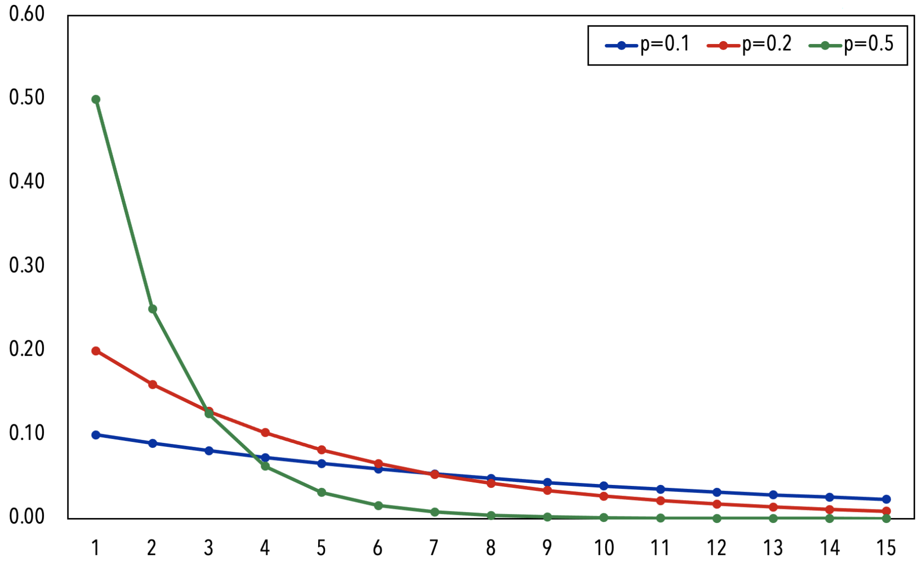

**Figure 5:** Comparison of geometric distribution of different parameters.