AP Stats Review

DRAFT – Not Final

Rafael A. Irizarry

April 9, 2025

Table of contents

Preface

These are draft notes intended to support review for AP Statistics. They are a work in progress and may change frequently as material is added or revised. They have not been fully proofread and are provided *as is*, without any guarantees regarding completeness or accuracy.

1 One varialbe data

1.1 The normal distribution

The normal distribution, also known as the bell curve and as the Gaussian distribution, is one of the most famous mathematical concepts in history.

One reason for this is that approximately normal distributions occur in many situations, including gambling winnings, heights, weights, blood pressure, standardized test scores, and experimental measurement errors.

This is not for discrete random variables but **continuous random variables**. If X has a normal pdf it can take on any continuous value. Therefore $Pr(X = x)$ does not make sense anymore.

The normal distribution is defined with a mathematical formula. For any interval (a, b), the proportion of values in that interval can be computed using this formula:

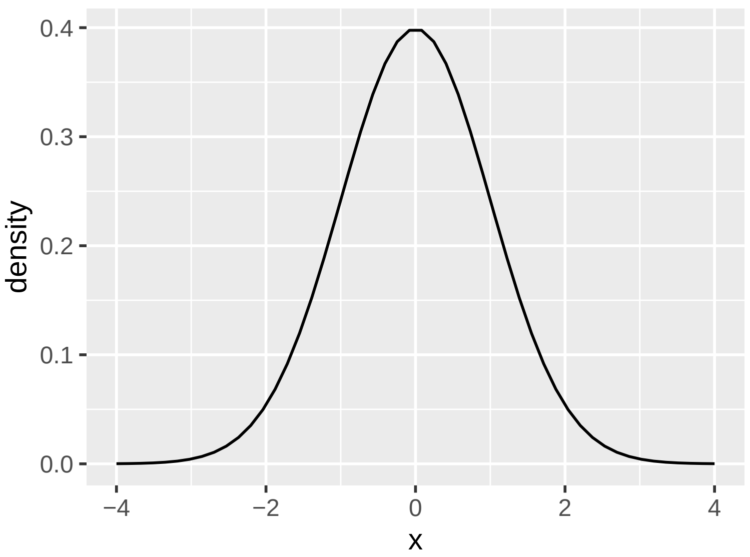
$$\mathrm{Pr}(a < x \leq b) = \int_a^b \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx$$

You don't need to memorize the formula.

The most important characteristics is that it is completely defined by just two parameters: μ and σ. The rest of the symbols in the formula represent the interval ends, a and b, and known mathematical constants π and e.

These two parameters, μ and σ, are referred to as the *mean* and the *standard deviation* (SD) of the distribution, respectively.

The distribution is symmetric, centered at μ, and most values (about 95%) are within 2σ from μ. Here is what the normal distribution looks like when the $\mu = 0$ and $\sigma = 1$:



If $\mu = 0$ and $\sigma = 1$ it is called **standard**.

Memorize the following: For a standard normal random variable Z:

- 68% are between -1 and 1
- 95% are between -2 and 2
- 99.7% between -3 and 3

You can use the `normalcdf` function to obtain these probabilities

- $\mathrm{P}(a < Z < b)$ is `normalcdf(a,b,0,1)`, 0 and 1 are mean and SD, respectively
- $\mathrm{P}(Z < a)$ is `normalcdf(-1E99,a,0,1)`. -1E99 is $-\infty$

Also memorize:

- If Z is normal and a is a constant aZ is normal.
- If Z is normal and b is a constant $Z + b$ is normal.
- If Z is normal and a, b are a constant $aZ + b$ is normal.
- If Z and W are normal and $Z + W$ is normal.

2 Random Variables and Probability Distributions

2.1 Discrete random variables

Discrete random variable

X can have different outcomes, each one with a probability. Discrete means the possible outcomes are finite.

Probability density function (pdf)

Defines the probability $\mathrm{P}(X = k)$ for each outcome k
The exam often uses short hand $p_k = \mathrm{P}(X = k)$

Note

Later we will need the *cumulative distribution function* which is simply defined by

$$\begin{aligned} F(a) &= \mathrm{P}(X <= a) \\ &= \sum_{x_i \leq a} \mathrm{P}(X = x_i) \end{aligned}$$

Interpretation $F(a)$ tells us the probability of X being less than a for any a.

Example 1: Fair coin

Define $X = 0$ for tails and $X = 1$ for heads

$$\begin{aligned} \mathrm{P}(X = 0) = 1/2 \\ \mathrm{P}(X = 1) = 1/2 \end{aligned}$$

Example 2: A die

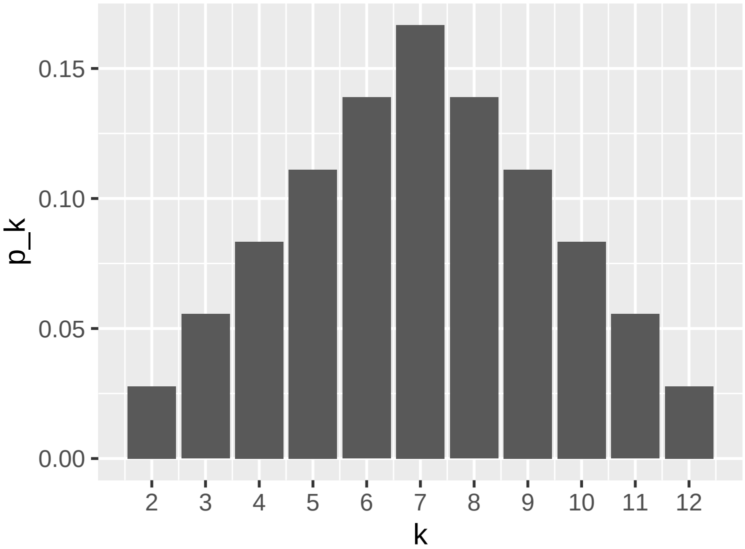
$$
\begin{gathered}
\mathrm{P}(X = 1) = 1/6 \\
\mathrm{P}(X = 2) = 1/6 \\
\vdots \\
\mathrm{P}(X = 6) = 1/6
\end{gathered}
$$

Example 3: Sum of two dice

$$
\begin{gathered}
\mathrm{P}(X = 2) = 1/36, \\
\mathrm{P}(X = 3) = 2/36 \\
\vdots \\
\mathrm{P}(X = 7) = 1/6 \\
\vdots \\
\mathrm{P}(X = 12) = 1/36
\end{gathered}
$$

The pdf is often shown as a table or a graph. For the graph we simply plot a bar for each k going up to p_k.

Here is th pdf for the sum of two dice:



2.2 Mean and standard deviation of a random variable

The mean and standard deviation of the the distribution of a random variable X are referred as the mean and standard deviation of X.

Mean

$$\mu_X = \sum_{i=1}^{n} x_i \, \mathrm{P}(X = x_i)$$

With

$$x_1, \dots, x_n$$

all the possible outcomes.

For any random variable $*$ we use the symbol μ_* to represent it's mean. It can be any random variable and we don't always use the name X.

The standard deviation (SD)

The standard de deviation of the distribution of a random variable X is defined as:

$$\sigma_X = \sqrt{\sum_{i=1}^{n} (x_i - \mu_X)^2 \, \mathrm{P}(X = x_i)}$$

We use σ_* just like μ_*

You can think of this as the typical distance you see X from the mean μ_X. For the heads or tails this is 1/2 since both 0 and 1 are1/2 from the mean 1/2.

ℹ Note

We sometimes say *the standard error of* X to mean *the standard deviation of the distribution of a random variable* X.

The variance

The variance is simply the standard deviation squared σ_X^2.

- We define the variance because mathematical calculations are easier without the square root. Other than that we always use standard deviation.
- The standard deviation has the same units as X. The variance has units squared which has no interpretation. What is kilograms squared or dollars squared?

2.3 Bernoulli trials

- A super common example of a useful random variable are Bernoulli trials.
- These are either a 0 (failure) or 1 (success) and each trial is independent of of others.

Bernoulli trial definition

- X is either 1 (success) or failure (0).
- Completely defined by the probability of success: $\mathrm{P}(X = 1) = p$.
- The probability of failure is simply $1 - p$, sometimes called q.

Bernoulli trials are popular because we can use them to count random things: number of heads when we toss coins, number of lottery winners, number of defective light bulbs made in a day by a factory, number of patients that got cured by a drug, number of COVID-19 hospitalizations in a day, and so on.

Bernoulli trial mean and SD

If X is a Bernoulli trial:

- $\mu_X = p$
- $\sigma_X = \sqrt{p(1-p)}$

You need to memorize this but here is the derivation

$$\begin{aligned} \mu_X &= 0 \times \mathrm{P}(X = 0) + 1 \times \mathrm{P}(X = 1) \\ &= p \end{aligned}$$

and

$$\begin{aligned} \sigma_X^2 &= (0-p)^2(1-p) + (1-p)^2 p \\ &= (1-p)p(p + 1 - p) \\ &= p(1-p) \end{aligned}$$

Examples

- Tossing coins, $p = 0.5$
- Steph Curry free throws, $p = 0.9$
- Lottery winners, $p < 10^{-6}$
- Celtics win a game in NBA finals $p = ?$

2.4 Combining, shifting, and scaling random variables

Mean of linear combinations

Need to memorize these (they are intuitive). If X and Y random variables and a is a constant:

- $\mu_{X+Y} = \mu_X + \mu_Y$
- $\mu_{X+a} = \mu_X + a$
- $\mu_{aX} = a\mu_X$

Example

If X and Y are two random variables, what is μ_{X-Y}?

$$\begin{aligned}\mu_{X-Y} &= \mu_X + \mu_{-Y} \\ &= \mu_X + -1\mu_Y \\ &= \mu_X - \mu_Y\end{aligned}$$

SD of linear combinations

For these we use the variance. But you can take square root at the end.

- $\sigma^2_{X+a} = \sigma^2_X$: shifting does not change variability.
- $\sigma^2_{aX} = a^2\sigma^2_X \implies \sigma_{aX} = |a|\sigma_X$: change of scale also scales measure of variability.
- **If** X **and** Y are independent, $\sigma^2_{X+Y} = \sigma^2_X + \sigma^2_Y \implies \sigma_{X+Y} = \sqrt{\sigma^2_X + \sigma^2_Y}$: Adding two things that vary, varies more.

Example

If X and Y are two **independent** random variables, what is σ_{X-Y}?

$$\begin{aligned}\sigma^2_{X-Y} &= \sigma^2_X + \sigma^2_{-Y} \\ &= \sigma^2_X + (-1)^2\sigma_Y \\ &= \sigma^2_X + \sigma^2_Y\end{aligned}$$

Which implies

$$\sigma_{X-Y} = \sqrt{\sigma^2_X + \sigma^2_{-Y}}$$

Interpretation: Subtracting two variables that vary independently has more variability than each.

2.5 Binomial distribution

Another popular random variable is the sum of Bernoulli trials.

$$S = \sum_{i=1}^{n} X_i$$

It tells us the number of successes and it is also a random variable.

Examples:

- Number of heads if I toss coins
- Number of free throws curry makes

Example

What is μ_S?

$$\begin{aligned} \mu_S &= \mu_{X_1+\cdots+X_n} \\ &= \mu_{X_1} + \cdots + \mu_{X_n} \\ &= np \end{aligned}$$

What is σ_S?

$$\begin{aligned} \sigma_S^2 &= \sigma^2_{X_1+\cdots+X_n} \\ &= \sigma^2_{X_1} + \cdots + \sigma^2_{X_n} \\ &= np(1-p) \end{aligned}$$

This implies

$$\sigma_S = \sqrt{np(1-p)}$$

Binomial pdf

We can compute the pdf for the sum of n trials:

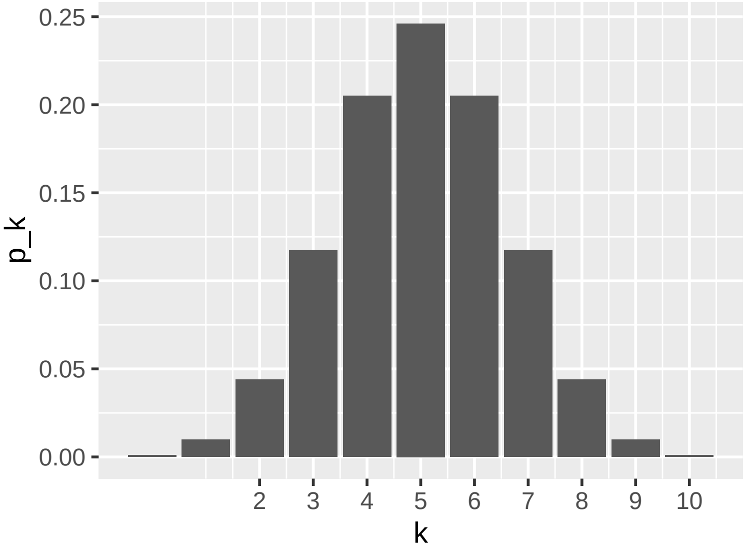
$$\mathrm{P}(S = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

This is called the binomial distribution and can be computed in AP-test with `binompdf(n, p, k)` and the CDF with `binomcdf(n,p,k)`

The CDF is useful for answering questions such as "what is the chance that we see 3 heads or less?" or "what is the chance we see 4,5,6 heads?"

Example

pdf of the number of heads when tossing 10 coins:



2.6 Geometric distribution

It is also common to ask how many trials do I need to see a success. For example, how many free throws will Curry take until he misses.

Geometric distribution

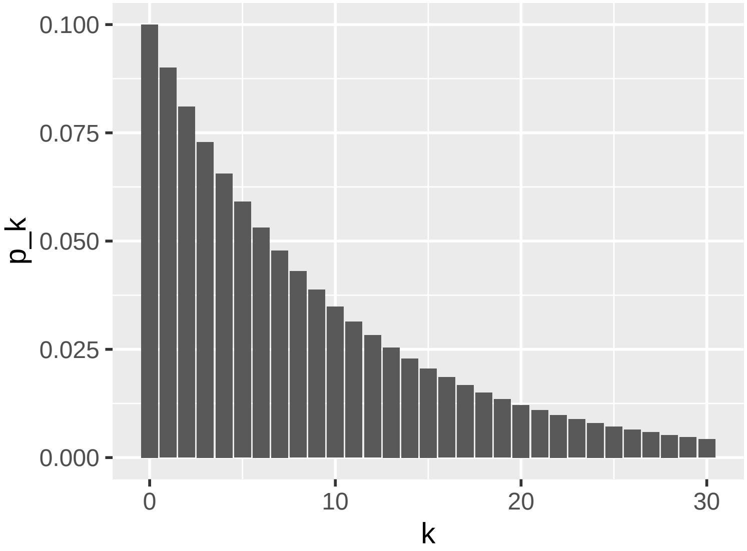
Define a random variable as X=number of trials if we stop after the first success.
It is not hard to see that this is:

$$\mathrm{P}(X = k) = (1 - p)^k p$$

This is called the Geometric distribution, defined for $k = 1, 2, \dots, \infty$.

Example: number of free thows before Curry misses.

Here miss is the success we are waiting for so $p = 0.1$



We use this to calculate, for example, that the chance of seeing 10 or more free throws in a row to start the game is `1 - geomcdf(10, .1)` = 0.3138106

2.7 Continuous random variables

Continuous random variables

Some random variables are continuous. Height, weight, and temperature are examples.

CDF

A continuous random variable X can take an infinity number of values x so it does not make sense to write:

$$\mathrm{P}(X = x)$$

Instead we define the cumulative distribution function as

$$\mathrm{F(a)} = \mathrm{P}(X < a)$$

We can then define the *probability density function* $f(x)$ so that

$$ F(a) = _{∞}^a f(x),dx

$$ :::{.callout-important} We use continuous distribution to approximate discrete ones. We will use

- normal distribution
- t distribution
- Chi-square distribution

In the test you either use a function on the calculator or they provide a look up table for the cdf $F(a)$.

:::

2.8 Approximation to Normal

When the number of trials is large binomial is very well approximated by the normal distribution.

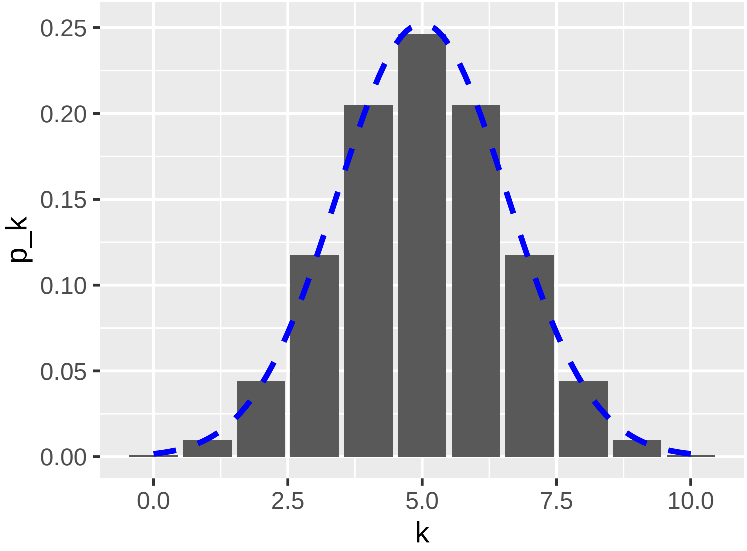
Define

$$Z = \frac{S - np}{\sqrt{np(1-p)}}$$

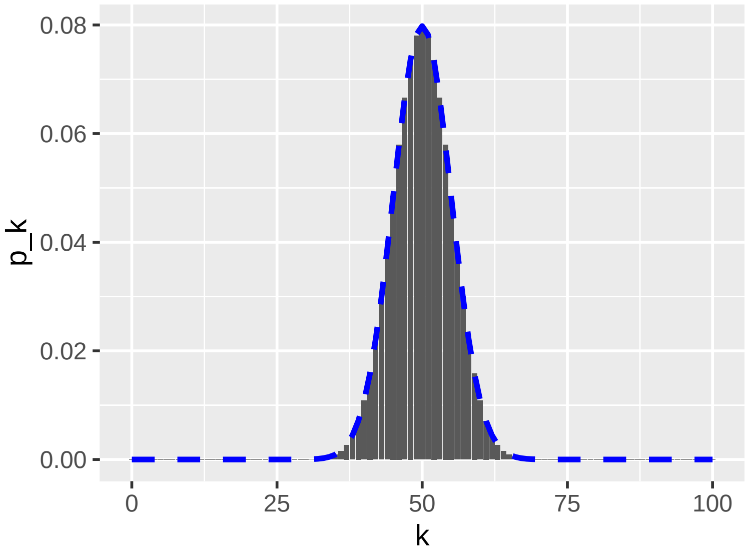
Then Z is approximated by standard normal

Here is a $n = 10, p = 0.5$ binomial with a normal with mean $np = 5$ and standard deviation $\sqrt{np(1-p)} \approx 1.6$ added in blue

```
Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
i Please use `linewidth` instead.
```



Here it is for 100. In this case $np = 50$ and $\sqrt{np(1-p)} = 5$



Example

If I toss 100 coins, what is the probability that I see between 45 and 55 heads?
We can use the binomial to answer this exactly `binomcdf(100, 0.5, 55) - binomcdf(100, 0.5, 44)` which is 0.728747.
But we can also use the normal distribution:

$$\begin{aligned}
\mathrm{P}(45 \leq S \leq 55) &= \mathrm{P}(44.5 < S < 55.5) \\
&= \mathrm{P}(44.5 - 50 < S - 50 < 55.5 - 50) \\
&= \mathrm{P}\left(\frac{44.5 - 50}{\sqrt{100 \times 0.5 \times 0.5}} < \frac{S - 50}{\sqrt{100 \times 0.5 \times 0.5}} < \frac{55.5 - 50}{\sqrt{100 \times 0.5 \times 0.5}}\right) \\
&= \mathrm{P}(-1.1 < Z < 1.1)
\end{aligned}$$

We use `normalcdf(-1.1 1.1, 0, 1)` = 0.7286679
Which is almost identical to the binomial result.

ℹ Note

Important to understand why we to the first $\mathrm{P}(45 \leq S \leq 55) = \mathrm{P}(44.5 < S < 55.5)$. The normal distribution is continuous so it can't be equal to anything. So we do the adjustment to make sure we include 45 and 55 in the approximation.

3 Sampling distributions

We want to learn about populations from samples.

3.1 Population and parameters

Population

The population is defined by the list of numbers $x_1, x_2, \ldots, x_n$.
The xs are **not random**.
We can't see the entire population but we want to learn about it.

Example 1: Trump voters

The population is the people who will vote on election day. Trump voters get a 1 and others get a 0. So all the xs are either 0 or 1.

Example 2: High school SAT scores

The population are all the students that took SAT. The xs are the scores for each student.

Population parameters

The population parameters are summaries of the $x_1, x_2, \ldots, x_n$ we are interested in.
In the AP test we almost always care about the **population mean**:

$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Most of this chapter is about *estimating* the population mean μ.
Another parameter we will need is the **population standard deviation**:

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2}$$

Population proportion

When the xs are 0s or 1s, then the population mean is equivalent to the proportion of 1s.
In this case we use the symbol p instead of μ.
And the standard deviation can be shown to be $\sqrt{p(1-p)}$

3.2 The sample average

The strategy to estimate the population parameter is to take a random sample: we can't examine all the sample so we examine a much smaller subset.

We learn that we can learn a lot about population parameters from samples.

By far the most common example is using a *sample average* to estimate a *population mean.*

A sample

- A sample are the resulting observed values we obtain when picking individuals at random from the population.
- We represent them with capital letters because they are **random variables**:

$$X_1, \dots, X_N$$

The sample size

N is called the sample size.
Do not confuse it with the number of individuals in the population n.
In the election poll example n is over 100 million while a typical sample size N is 1,000 or less.

The sample average

- A sample are the resulting observed values we obtain when picking individuals at random from the population.
- We represent them with capital letters because they are **random variables**:

$$X_1, \dots, X_N$$

* N is called the sample size.

3.3 Central Limit Theorem

tldr: The distribution of the sample average is approximated by a normal distribution when the sample size is large.

Central Limit Theorem (CLT)

- If X_1, ...X_N are random variables that are independent and have the same distribution, the sum $\sum_{i=1}^{N} X_i$ gets closer and closer to being normally distributed when N gets very large.
- Because dividing a normal random variable by a constant is still normal, the CLT applies to the average $\frac{1}{N}\sum_{i=1}^{N} X_i$ as well.
- **Rule of thumb** $N \geq 30$ is considered large enough.

3.4 Proportions

A very common application of statistics is estimating a population proportion

Examples:

- Proportion of voters voting for trump.
- Proportion of patients that a drug cures.
- Proportion of adults with a job.

We want to to estimate p, the population parameter.

Note that:

- Each X in the sample is a Bernoulli trial because $\mathrm{P}(X = 1) = p$.
- This implies that for all i, $\mu_{X_i} = p$ and $\sigma_{X_i} = \sqrt{p(1-p)}$
- Because we sample with replacement the Xs are independent.

Mean and SD of sample proportion

The sample proportion is

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} X_i$$

Using what we have learned about mean and SD of combinations and re-scaling we have:

- $\mu_{\hat{p}} = p$
- $\sigma_{\hat{p}} = \frac{\sigma_X}{\sqrt{N}} = \frac{\sqrt{p(1-p)}}{\sqrt{N}}$

Distribution of sample proportion

- $\hat{p}$ is a sum of Bernoulli trials divided by a constant. So we could use the Binomial distribution to compute $\mathrm{P}(\hat{p} = k/N)$
- However, in the exam they want you to use the CLT.
- $\hat{p}$ is approximated by normal distribution with mean p and SD $\sqrt{p(1-p)/n}$

Example

If I take a poll if 1000 people to get an idea of how many people are voting for Trump, what is the chance that my sample proportion $\hat{p} = 0.45$ is within 1% of the actual proportion?
We are asking $\mathrm{P}(|\hat{p} - p| < 0.01)$
Let's figure it out:

$$\begin{aligned}
\mathrm{P}(|\hat{p} - p| < 0.01) &= \mathrm{P}(-0.01 < \hat{p} - p < 0.01) \\
&= \mathrm{P}\left(\frac{-0.01}{\sqrt{\frac{p(1-p)}{N}}} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{N}}} < \frac{0.01}{\sqrt{\frac{p(1-p)}{N}}}\right) \\
&= \mathrm{P}\left(\sqrt{1000}\frac{-0.01}{\sqrt{p(1-p)}} < Z < \sqrt{1000}\frac{0.01}{\sqrt{p(1-p)}}\right)
\end{aligned}$$

I don't know p but in the exam they want you to stick in $\hat{p}$ for the SD calculation. So $\sqrt{1000/(0.45 \times .55)} \approx 63.5$
so we have $\mathrm{P}(0.645 < Z < 0.645)$ or `normcdf(-0.645, 0.645, 0, 1)` which is r `pnorm(0.645)-pnorm(-0.645)`

Tip

In the exam compute the standard deviation $\sqrt{p(1-p)/N}$ first and stick that in the calculations instead of the formula.

3.5 Means

Another common application of statistics is estimating a population mean

Examples:

- What is the average SAT score in a high school?
- What is the average blood pressure for people taking a drug?

We want to to estimate μ, the population parameter.

Note that

- Each X in the sample has the same distribution $\mathrm{P}(X = x_i) = 1/n$ for all i.
- This implies that for all i, $\mu_{X_i} = \mu$ and $\sigma_{X_i} = \sigma$
- Because we sample with replacement the Xs are independent.

Mean and SD of sample average

The sample average is

$$\bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i$$

Using what we have learned about mean and SD of combinations and re-scaling we have:

- $\mu_{\bar{X}} = \mu$
- $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{N}}$

Distribution of sample average

CLT tells us that $\bar{X}$ is approximated by a normal distribution with mean μ and SD $\sqrt{\sigma/n}$

Sample standard deviation

If I want to make probability calculations I need to know $\sigma_{\bar{X}}$, but I don't know *sigma*.
For proportions we used $\sqrt{\widehat{p}(1-\widehat{p})}$ as an approximation of the standard deviation $\sqrt{p(1-p)}$
But when sample means are not based on Bernoulli trials, we can't do that.
Instead we use the sample standard deviation.

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

4 Proportions

4.1 Confidence intrval

Task:

- We take a sample of 1,000 voters.
- 45% of our respondents say they will vote for trump.
- Provide an interval with 95% of containing the true proportion p

Solution:

Our proportion $\hat{p}$ follows a normal distribution with mean p and standard deviation $\sqrt{p(1-p)/N} \approx \sqrt{0.45 \times 0.55/1000} \approx 0.016$

Lets consider symmetric intervals with: $[\hat{p} - B, \hat{p} + B]$

We want to find a *margin of error* MOE such that:

$$\mathrm{P}(p \in [\hat{p} - \mathrm{MOE}, \hat{p} + \mathrm{MOE}]) = 0.95$$

We accomplish it by setting $\mathrm{MOE} = 2\sigma_{\hat{p}} \approx \sqrt{\hat{p}(1-\hat{p})}/\sqrt{N} \approx 0.03$

$$\begin{aligned}
\mathrm{P}(p \in [\hat{p} - 2\sigma_{\hat{p}}, \hat{p} + 2\sigma_{\hat{p}}]) &= \mathrm{P}(\hat{p} - 2\sigma_{\hat{p}} < p < \hat{p} + 2\sigma_{\hat{p}}]) \\
&= \mathrm{P}(-2\sigma_{\hat{p}} < \hat{p} - p < 2\sigma_{\hat{p}}]) \\
&= \mathrm{P}\left(-2 < \frac{\hat{p} - p}{\sigma_{\hat{p}}} < 2\right) \\
&= \mathrm{P}(-2 < Z < 2) \\
&= 0.95
\end{aligned}$$

So our interval is 0.45 ± 0.03.

- If we want to be 99.7% sure we can use 3 instead of 2.
- If we want to be 68% sure we can use 1 instead of 2.

We refer to the 0.03 as the *margin of error* (MOE).

4.2 Critical values

We already knew that using 2 would give us a 95% confidence interval.

But what if we didn't know? Or if we wanted a 99% confidence interval?

The function `invNorm` will do this for us. The impute is the area to the left.

So to obtain 95% we need 0.5% to the left and 0..5% to the right.

We use `invNorm(0.995)` which gives us `r qnorm(0.995)`

So we multiply by 2.57 not 2 to get a 99% confidence interval.

Note that to get exactly 95% we actually use `invNorm(0.975)` which is `r qnorm(0.975)`, a little bit less than 2. In some books you will see 1.96 instead of 2.

4.3 p-values

- We want to know if a coin in biased.
- We toss it 100 times and observe 60% heads.

Is it biased or can this happen by chance?

Let's compute the probability of seeing $\hat{p} = 0.6$ or more extreme.

Note that 0.4 is as extreme: we usually permit both directions.

Null hypothesis: It is fair or $p = 0.5$

We will reject if the *p-value* is 0.05 or smaller.

The p-value is the probability observing something as extreme as we did when the null hypothesis holds

$$\begin{aligned}
\mathrm{P}(|\hat{p} - p| \geq 0.1) &= 1 - \mathrm{P}(|\hat{p} - p| < 0.1) \\
&= 1 - \mathrm{P}\left(\left|\frac{\hat{p} - p}{\sigma_{\hat{p}}}\right| < \frac{0.1}{\sigma_{\hat{p}}}\right) \\
&= 1 - \mathrm{P}\left(|Z| < \frac{0.1}{\sigma_{\hat{p}}}\right)
\end{aligned}$$

When the null hypothesis holds, $\sigma_{\hat{p}} = \sqrt{0.5 \times 0.5/100} = 0.05$

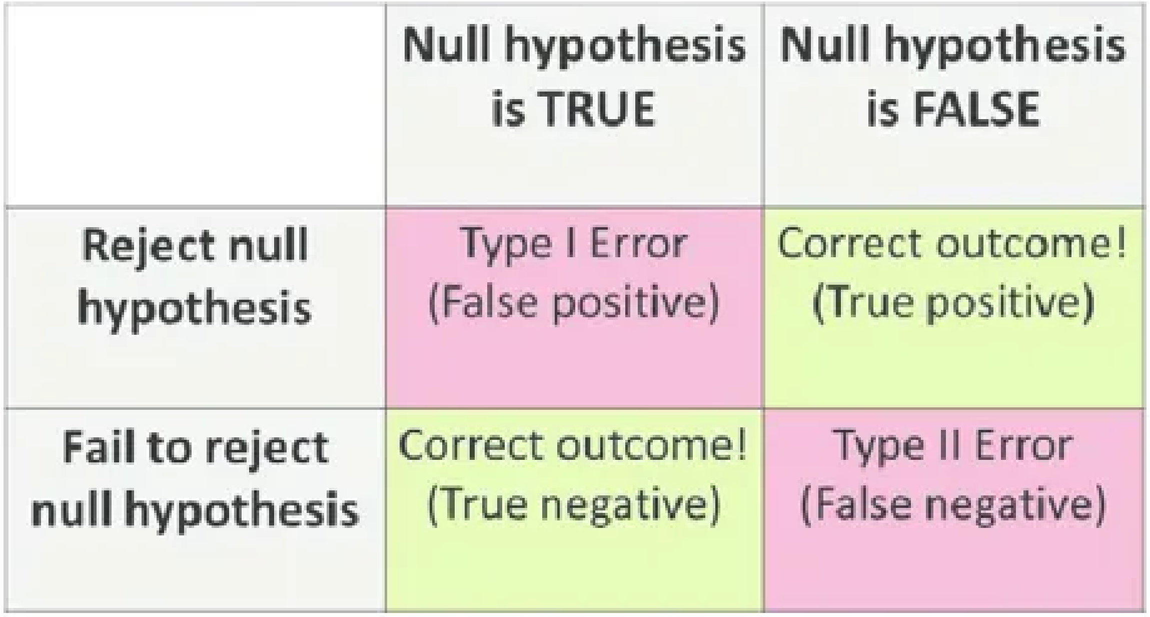
So the p-value is $1 - \mathrm{P}(|Z| < 0.1/0.05 = 2)$ which is a bit less than 0.05

We reject the null hypothesis.

Type of errors

- Type I error is rejecting the null hypothesis when it is true. Example say the coin is biased when it was fair.
- Type II error is failing to reject the null hypothesis when it is not true. Example saying the coin is fair when it was biased.
- Power is the 1 - probability of Type II error.

To help remember:



	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)

4.4 Difference of two proportion

4.4.1 Confidence interval

- Does drug work better than placebo?
- The proportion of the populations are p_1 and p_2.
- For both placebo and drugged populations we obtain sample means $\hat{p}_1$ and $\hat{p}_2$
- The sample sizes are N_1 and N_2
- Provide a 95% confidence interval

We know the difference $\hat{p}_1 - \hat{p}_2$ had the following mean and SD:

- $\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2$
- $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{N_1} + \frac{p_2(1-p_2)}{N_2}}$

To construct a 95% confidence interval we use $\hat{p}_1 - \hat{p}_2 \pm \sigma_{\hat{p}_1-\hat{p}_2}$

As before we estimate

$$\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{N_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{N_2}}$$

4.4.2 p-value

- Suppose we have sample sizes of 25 and 100 for the the drug and placebo group respectively and
- we observe $\hat{p}_1 = 0.25$ and $\hat{p}_2 = 0.15$
- The null is that there is no difference so $p_1 - p_2$ or $p_1 = p_2 = p$

Under the null hypothesis we have

- $\mu_{\hat{p}_1-\hat{p}_2} = 0$
- $\sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p(1-p)}{N_1} + \frac{p(1-p)}{N_2}}$

To compute the p-value we need an estimate for $\sigma_{\hat{p}_1-\hat{p}_2}$ which depends on p

We estimate p with pooled data:

$$\frac{N_1\hat{p}_1 + N_2\hat{p}_2}{N_1 + N_2} = 0.17$$

which means $\sigma_{\hat{p}_1-\hat{p}_2} \approx \sqrt{(0.17 \times 0.83)(1/25 + 1/100)} \approx 0.08$

With this we can compute

$$\begin{aligned} \mathrm{P}(|\hat{p}_1 - \hat{p}_2| \geq 0.1) &= 1 - \mathrm{P}(|Z| < 0.1/\sigma_{\hat{p}_1-\hat{p}_2}) \\ &= 1 - \mathrm{P}(|Z| < 0.1/0.08) \\ &\approx 0.20 \end{aligned}$$

We do not reject.

4.4.3 Confidence interval and p-value connection

You can do the math and see that **if a 95% confidence interval does not include the null hypothesis mean, then a p-value will be less than 0.05**

The math:

If the null hypothesis says the mean is p and the observed $\hat{p}$ resulted in a p-value less than 0.05, we know:

$$\left|\frac{\hat{p} - p}{\sigma_{\hat{p}}}\right| > 2$$

This implies that either

$$p > \hat{p} + 2\sigma_{\hat{p}} \text{ or } p < \hat{p} - 2\sigma_{\hat{p}}$$

5 Means

5.1 Confidence intrval

Task:

- We take a sample of 36 student SAT scores.
- We observe a sample average of $\bar{X} = 1100$ and a sample standard deviation $s = 204$
- Provide an interval with 95% of containing the high school population average μ.

Solution:

The sample average $\bar{X}$ follows a normal distribution with mean μ and standard deviation $\sigma/\sqrt{N} \approx 204/6 = 34$

As with proportions we have

$$\begin{aligned} \mathrm{P}(\mu \in [\bar{X} - 2\sigma_{\bar{X}}, \bar{X} + 2\sigma_{\bar{X}}]) &= \mathrm{P}(\bar{X} - 2\sigma_{\bar{X}} < \mu < \bar{X} + 2\sigma_{\bar{X}}]) \\ &= \mathrm{P}(-2\sigma_{\bar{X}} < \bar{X} - \mu < 2\sigma_{\bar{X}}) \\ &= \mathrm{P}\left(-2 < \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < 2\right) \\ &= \mathrm{P}(-2 < Z < 2) \\ &= 0.95 \end{aligned}$$

So our interval is 1100 ± 68

5.2 t-test

When $N < 30$ we can't use CLT.

So what is the distribution of $\bar{X}$?

If the population values are also approximately normal, as they are for SAT scores, then

$$t = \frac{\bar{X} - \mu}{s/\sqrt{N}}$$

Follows a t-distribution with $N-1$ degrees of freedom.

Example

Let's repeat the above example but $N = 15$
All we have to do now is use the cutoff that gives us 0.95 for a t-distribution with 14 degrees of freedom.
You can use `invT` with area to the left and degrees of freedom.
So instead of 2 we use `invT(0.975, 14)` which is `r qt(0.975, 14)`, a little bit bigger than 2.
We make our confidence interval

$$1100 \pm 2.14 \times 34$$

or

$$1100 \pm 73$$

5.3 Difference of two means

5.3.1 Confidence interval

- Are the mean SAT scores in two high schools different?
- The sample averages are $\bar{X}_1 = 1200$ and $\bar{X}_2 = 1100$ and the sample standard deviations are $s_1 = 200$ and $s_2 = 180$
- The sample sizes are $N_1 = 30$ and $N_2 = 35$
- Provide a 95% confidence interval

We know the difference $\bar{X}_1 - \bar{X}_2$ had the following mean and SD:

- $\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - mu_2$
- $\sigma_{\bar{X}_1 - \bar{X}_2} = \sqrt{\frac{\sigma_1^2}{N_1} + \frac{\sigma_2^2}{N_2}}$

To construct a 95% confidence interval we use $\bar{X}_1 - \bar{X}_2 \pm \sigma_{\bar{X}_1 - \bar{X}_2}$

We approximate σ_1 and σ_2 with s_1 and s_2

$$\sigma_{\bar{X}_1-\bar{X}_2} \approx \sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}$$

The sample sizes are large enough that we can use CLT so the confidence interval is

$$\bar{X}_1 - \bar{X}_2 \pm 2\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}}$$

$$\bar{X}_1 - \bar{X}_2 = 100$$

and

$$2\sqrt{\frac{s_1^2}{N_1}+\frac{s_2^2}{N_2}} = 2\sqrt{200^2/30+180^2/35} \approx 48$$

So the confidence interval is

$$100 \pm 48$$

5.3.2 p-value

Is the difference we saw significant?

We already computed $\sigma_{\bar{X}_1-\bar{X}_2} \approx 24$

$$\mathrm{P}(|\bar{X}_1 - \bar{X}_2| > 100) = \mathrm{P}(|Z| > 100/47) \approx 0.03$$

We reject.

6 Goodness of fit

Example:

Are all color Skittles equally likely?
Here is the dat?

Color	Observed Count
Red	20
Yellow	25
Green	15
Purple	18
Orange	22
Total	**100**

We see more yellow and less green?
Can this happen by chance?

Chi-square test

This is a **goodness of fit** test.
If we have k categories each with p_i, $i = 1, \dots, k$ and observe N outcomes, we expect to seen $E_i = Np_i$ of each.
If we observe O_i for categories $i = 1, \dots, k$ then the distribution of

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

follows what is called a chi-square distribution with $k - 1$ degrees of freedom.

Example:

Getting back to our example, if equally likely we have

$$E_i = 100/5 \approx 20$$

for all categories.
We can compute the χ^2 stat by

$$\left((20-20)^2+(25-20)^2+(15-20)^2+(18-20)^2+(22-20)^2\right)/20 = 2.8$$

We can look up this probability for a χ^2 with 4 degrees of freedom and see that the p-value is 0.6.
So can easilty see this by chance.

Example

Is promotion status **independent** of gender, or is there evidence of **gender bias**?
Observed Data:

Gender	Promoted	Not Promoted	Total
Men	45	55	100
Women	30	70	100
Total	**75**	**125**	**200**

Calculate expected counts under the assumption of independence:

$$E_{ij} = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}$$

Gender	Promoted	Not Promoted
Men	$\frac{100\times75}{200} = 37.5$	$\frac{100\times125}{200} = 62.5$
Women	37.5	62.5

Compute the chi-square test statistic:

$$\chi^2 = \sum \frac{(O-E)^2}{E} = \frac{(45-37.5)^2}{37.5} + \frac{(55-62.5)^2}{62.5} + \frac{(30-37.5)^2}{37.5} + \frac{(70-62.5)^2}{62.5}$$

$$= \frac{56.25}{37.5} + \frac{56.25}{62.5} + \frac{56.25}{37.5} + \frac{56.25}{62.5} = 1.5 + 0.9 + 1.5 + 0.9 = \mathbf{4.8}$$

This has 1 degree of freedom.
Use the chi-square cumulative distribution function:

$$P(\chi^2 \geq 4.8) \approx 0.028$$

Since the p-value is approximately 0.028, which is less than the typical significance level of 0.05, we **reject** H_0.

! Important

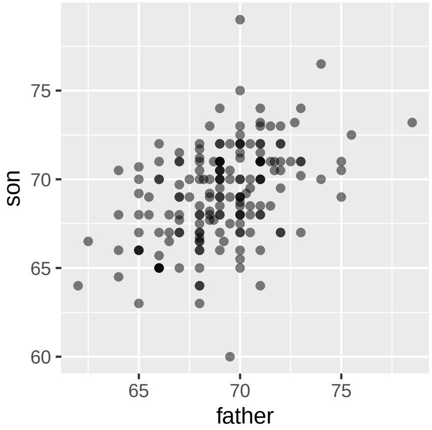
In the test there will be two types of problems related ti goodness of fit:

- Is the distribution of categorial data as expected (for example, Skittles)?
    - Several categories k
    - probabilities of each category given
    - degrees of freedom is $k-1$
- Are two variables independent (for example, gender bias)?
    - Two categories
    - proportions calculated are from data.
    - degrees of freedom is 1.

7 Slopes

Example is height hereditary?

How well can we predict a child's height based on the parents' height?
We can summarize the data with the two averages and two standard deviations. However, this summary fails to describe an important characteristic of the data: the trend that the taller the father, the taller the son.



The correlation coefficient

The correlation coefficient is defined for a list of pairs $(x_1, y_1), \dots, (x_n, y_n)$ as the average of the product of the standardized values:

$$\rho = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \mu_x}{\sigma_x} \right) \left(\frac{y_i - \mu_y}{\sigma_y} \right)$$

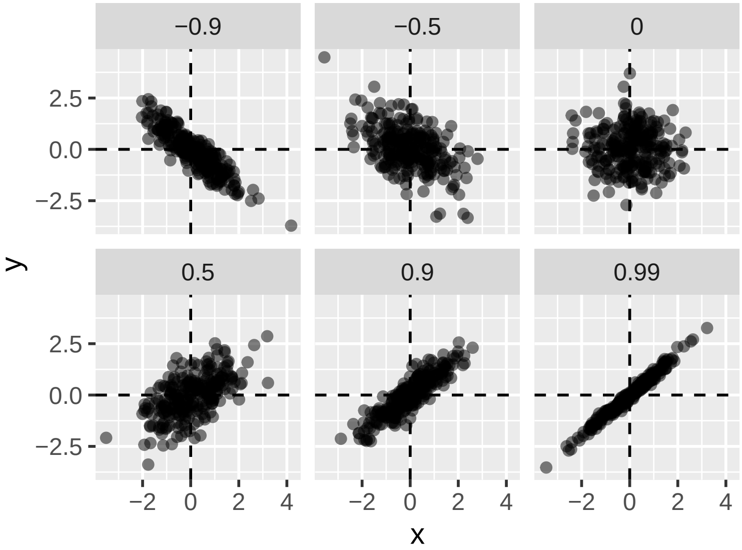
with μ_x, μ_y the averages of $x_1, \dots, x_n$ and $y_1, \dots, y_n$, respectively, and σ_x, σ_y the standard deviations.

To understand why this equation does in fact summarize how two variables move together, consider the i-th entry of x is $\left(\frac{x_i - \mu_x}{\sigma_x}\right)$ SDs away from the average. Similarly, the y_i that is paired with x_i, is $\left(\frac{y_1 - \mu_y}{\sigma_y}\right)$ SDs away from the average y. If x and y are unrelated, the product $\left(\frac{x_i - \mu_x}{\sigma_x}\right)\left(\frac{y_i - \mu_y}{\sigma_y}\right)$ will be positive ($+ \times +$ and $- \times -$) as often as negative $(+ \times -$

and $- \times +$) and will average out to about 0. This correlation is the average and therefore unrelated variables will have 0 correlation. If instead the quantities vary together, then we are averaging mostly positive products ($+ \times +$ and $- \times -$) and we get a positive correlation. If they vary in opposite directions, we get a negative correlation.

The correlation coefficient is always between -1 and 1.

To see what data looks like for different values of ρ, here are six examples of pairs with correlations ranging from -0.9 to 0.99:



Sample correlation

The ρ defined above is for a population $(x_1, y_1), \ldots, (x_n, y_n)$
If we have a sample $(X_1, Y_1), \ldots, (x_N, y_N)$we can estimate with the sample correlation:

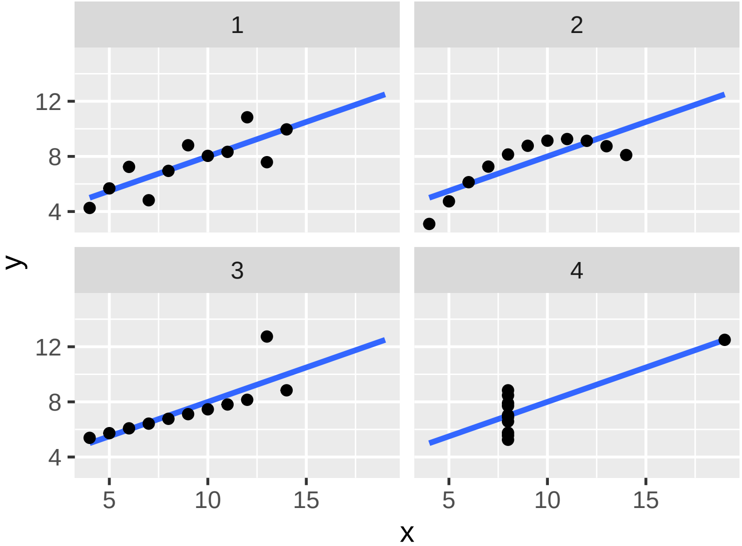
$$r = \frac{1}{N-1} \sum_{i=1}^{N} \left(\frac{X - \bar{X}}{s_X} \right) \left(\frac{Y - \bar{Y}}{s_Y} \right)$$

! Important

r is a random variable!

7.1 Correlation is not always a useful summary

Correlation is not always a good summary of the relationship between two variables. The following four artificial datasets, referred to as Anscombe's quartet, famously illustrate this point. All these pairs have a correlation of 0.82:



The regression line

If we are predicting a random variable Y knowing the value of another $X = x$ using a regression line, then we predict that for every standard deviation, σ_X, that x increases above the average μ_X, our prediction $\hat{Y}$ increase ρ standard deviations σ_Y above the average μ_Y with ρ the correlation between X and Y. The formula for the regression is therefore:

$$\left(\frac{\hat{Y} - \mu_Y}{\sigma_Y}\right) = \rho \left(\frac{x - \mu_X}{\sigma_X}\right)$$

We can rewrite it like this:

$$\hat{Y} = \mu_Y + \rho \frac{\sigma_Y}{\sigma_X}(x - \mu_X)$$

- The $\rho \frac{\sigma_Y}{\sigma_X}$ of the regression line is proportional to r.
- You can think of $\frac{\sigma_Y}{\sigma_X}$ as needed for unit conversion.

If there is perfect correlation, the regression line predicts an increase that is the same number of SDs. If there is 0 correlation, then we don't use x at all for the prediction and simply predict the average μ_Y. For values between 0 and 1, the prediction is somewhere in between. If the correlation is negative, we predict a reduction instead of an increase.

Note that if the correlation is positive and lower than 1, our prediction is closer, in standard units, to the average height than the value used to predict, x, is to the average of the xs. This is why we call it *regression*: the son regresses to the average height. In fact, the title of Galton's paper was: *Regression toward mediocrity in hereditary stature*. To add regression lines to plots, we will need the above formula in the form:

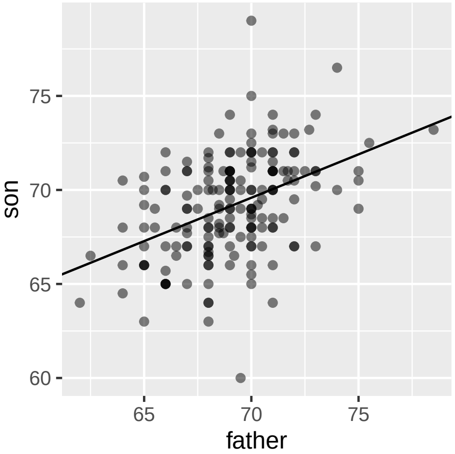
$$\hat{Y} = b + mx \text{ with slope } m = \rho \frac{\sigma_y}{\sigma_x} \text{ and intercept } b = \mu_y - m\mu_x$$

Estimating the regression line

The equation above is theoretical. The estimate using the data is

$$\hat{Y} = \bar{Y} + r\frac{s_Y}{s_X}(x - \bar{X})$$

Here we add the regression line to the original data:



The regression formula implies that if we first standardize the variables, that is subtract the average and divide by the standard deviation, then the regression line has intercept 0 and slope equal to the correlation ρ. You can make same plot, but using standard units like this: