Review of Probability and Statistics

ECON 4753 — University of Arkansas

Prof. Kyle Butts

1 — Single Random Variables

There are two key concepts to understand. First, we have an **experiment** which is the source of *randomness* in the world: e.g. flip a coin, roll a dice, play a basketball game, et cetera. The second definition is a **random variable**. A random variable assigns numerical values that is determined by an experiment (e.g. value of a die roll, assigning 1 to heads 0 to tails). For an experiment, there are many different random variables you could define. For a NBA game, random variables include the number of points scored; a variable who equals 1 if the home team won and 0 if the home team loses; the average height of players who play; etc.

We denote random variables by upper case letters, usually W, X, Y, and Z. The particular values that a random variable takes are denoted by the corresponding lower case letters w, x, y, and z.

For , think about a the coin-flipping experiment where you flip 10 coints. Let X denote the random variable counting the number of heads that land out of the 10 flips. It is important to note X is not associated with any particular value. You can do this experiment many different times and X represents *the process* of running the experiment, not the value of a particular trial. But, we know X will take on a value in the set $\{0, 1, 2, \ldots, 10\}$, say $x = 6$ for a particular trial.

Discrete Random Variables

When the random variable takes on **discrete** values (like X above), the probability density function (pdf) assigns probabilities for X obtaining any particular value. In the case of discrete values, this is sometimes called the probability mass function. Formally, say X can take n values. For all values, denoted by x_j, that the random variable can take, the pdf is defined as:

$$f_X(x_j) = \mathbb{P}(X = x_j) = p_j$$

Example 1.1. Say you flip a single coin and assign X to equal 1 if the coin lands on heads and 0 if it lands on tails. The pdf for this random variable is $f(1) = 1/2$ and $f(0) = 1/2$.

There are two rules that the pdf must follow:

1. For all values of x_j, $1 \geq f_X(x_j) \geq 0$
2. $\sum_j f_X(x_j) = 1$

In words, (1) says the probability X takes a value x_j has to be between 0 and 1, and (2) says the sum of the probabilities *over all possible values* must equal 1 (i.e. something must happen).

The **cumulative distribution function** (cdf) is similar to the pdf but asks about the probability that X takes a value less than or equal to some x:

$$F_X(x) = \mathbb{P}(X \leq x)$$

Note that this means CDF must be (weakly) increasing in x (e.g. $\mathbb{P}(X \leq 2)$ must be no larger than $P(X \leq 3)$). Also $0 \leq F_X(x) \leq 1$ must be true since that is true of probabilities.

Example 1.2. Take a random variable, Y with pdf of $p_1 = \mathbb{P}(Y = 1) = 0.2$, $p_2 = \mathbb{P}(Y = 2) = 0.5$, and $p_3 = \mathbb{P}(Y = 4) = 0.3$. Then, the CDF is given by

$$F_y(y) = \begin{cases} 0.0 & \text{when } y < 1, \\ 0.2 & \text{when } 1 \leq y < 2, \\ 0.7 & \text{when } 2 \leq y < 4, \\ 1.0 & \text{when } y \geq 4, \end{cases}$$

■

The probability density function is *a complete picture* about a random variable; but it can often be too much information (imagine telling someone the probabilities for X that can take the values 0 to 100.) While you could plot the pdf as a histogram, this still can be a lot of information. For that reason, we use **statistics** to 'summarize a random variable'.

Expectation

The most common statistic we use is the mean, or in fancy words the 'expectation'. A natural question we care about is 'what is the average value of some random variable X?' For this, we use the **expectation** which takes the average of the values that X can take weighted by how likely X is to take that value:

$$\mathbb{E}(X) \equiv \sum_{j=1}^{n} x_j \, \mathbb{P}(X = x_j) = \sum_{j=1} x_j p_j$$

Example 1.3. Take a random variable, Y with pdf of $p_1 = \mathbb{P}(Y = 1) = 0.2$, $p_2 = \mathbb{P}(Y = 2) = 0.5$, and $p_3 = \mathbb{P}(Y = 4) = 0.3$. Then, the expectation is given by

$$\mathbb{E}(Y) = 1 * 0.2 + 2 * 0.5 + 4 * 0.3 = 2.4$$

■

There are some properties of expectations that come out of the definitions of sums $\sum$:

1. For any constant c, $\mathbb{E}(c) = c$ (there's no randomness)
2. For any constant a, $\mathbb{E}(aX) = a\,\mathbb{E}(X)$
3. For any constant a and b, $\mathbb{E}(aX + bY) = a\,\mathbb{E}(X) + b\,\mathbb{E}(Y)$

Lastly, we will discuss transformations of random variables. So for , say you have a function g that takes the random variable, X, of the number of heads out of 10 and assign to sum final score, Y. That is, $Y = g(X)$. The average value of Y can be calculated using the pdf of X:

$$\mathbb{E}(Y) = \mathbb{E}(g(X)) = \sum_{j=1}^{n} p_j g(x_j)$$

Variance

The other most common statistic is the **variance** of a random variable. While the expectation tells you something about the 'average' value you would observe the random value taking, the variance tells you something about the varaibility of the random variable. That is, how much does the random variable X move around it's mean? A large variance means it moves around a lot.

The variance is given by

$$\text{var}\,(X) \equiv \mathbb{E}\big((X - \mathbb{E}(X))^2\big) = \sum_j p_j (x_j - \mathbb{E}(X))^2,$$

where the last equality comes from the $g(X)$ rule we laid out above.

The intuition around the variance is as follows. The value $x_j - \mathbb{E}(X)$ tells you how far a particular x value is from the expected value. If we just used the difference, then the positives (x_j above the mean) and the negatives (x_j below the mean) would cancel out which would make this a bad measure of variability. So, therefore, we square the term to make it positive everywhere.

Relatedly, the **standard deviation** is the square-root of the variance:

$$\text{sd}(X) \equiv \sqrt{\text{var}\,(X)}$$

In your own time, check the following using the definition of variance above:

1. For any constant c, we have $\text{var}\,(c) = 0$
2. For any constants a and b, $\text{var}\,(aX + b) = a^2\,\text{var}\,(X)$

Continuous Random Variables

Now we turn to **continuous** random variables where random variables can take values on (parts of) the real number line (e.g. height). All our definitions and intuition still apply, but in continuous land, we basically move from sums $\sum$ to integrals $\int$.

The probability density function is given by $f_X(x)$, but no longer represents the probability that X obtains the value x. This is because with a continuum of values, the probability you obtain any *particular* value becomes 0. E.g. what's the probability someone is *exactly* 5 ft. 8.124028 in.?[1] Instead, it is better to think of the density function as giving a 'relative likelihood' that you take a value "near x". More formally, the density function can tell you the probability you're in a range of values $[\underline{x}, \bar{x}]$:

$$\mathbb{P}(X \leq x) = \int_{\underline{x}}^{\bar{x}} f_X(x) dx,$$

1. As an aside, you might argue 'well with a ruler we would round these variables to the nearest tenth of an inch', so isn't that discrete? I guess you can make this argument and treat it as a discrete variable, but it is better to think of these things as continuous (later in the class it will make the math simpler, not harder!) and we will in this class.

This can be used for the CDF too:

$$\mathbb{P}(X \leq x) = \int_{-\inf}^{x} f_X(x)dx,$$

Similarly, expectations still take the form of "the value times the density", but we 'average' this using the integral:

$$\mathbb{E}(X) = \int x * f_X(x)dx.$$

All the properties listed above hold for continuous variables too (or mixtures of both).

Last, the variance is given by

$$\text{var}\,(X) = \int (x - \mathbb{E}(X))^2 f_X(x)dx$$

2 — Multiple Random Variables

In this class, we care about how variables relate to one another. For , do sales grow and shrink with a consumer's age?; is height an important predictor of basketball success?; is the sale associated with a large increase in sales?

For this, we would want to know about the **joint distribution** between two variables X and Y.

The **joint probability density function** is denoted by $f_{X,Y}(x, y)$. In the discrete case, it is the probability that $X = x$ *and* $Y = y$ *in the same trial*:

$$f_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y).$$

We can also think about **conditioning** on the value of one of the random variables. That is, take Y to be *fixed* to some value y. Then we could ask about the distribution of X *within trials where* $Y = y$.

$$f_{X|Y}(x|y) = \mathbb{P}(X = x \mid Y = y)$$

Note we use the | symbol to note 'conditioning' on some random variable's realization. The things on the right we are conditioning on. In our , we learn that $Y = y$ for that trial and then ask about

the (conditional) probability that $X = x$.

The **Bayes Rule** helps us translate between conditional pdfs and the joint pdfs:

$$f_{X|Y}(x|y) = f_{X,Y}(x, y)/f_Y(y).$$

So say we have two discrete random variables and we want to know probability $X = x$ conditional on $Y = y$. We can calculate it if we know (1) the probability that $Y = y$ and (2) the joint probability that $Y = y$ *and* $X = x$.

Covariance and Correlation

Similar to how we used statistics to summarize a single random variable, it is common to want to summarize how to variables are related to one another. For this we use the **covariance** or the **correlation** between two variables. They are very similar to one another.

The covariance looks like the variance of a random variable:

$$\text{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right]$$

The covariance intuitively measures whether X and Y move together. The covariance is positive if when X is above it's mean, Y also tends to be above it's mean and when X is below it's mean, Y also tends to be below it's mean. They "co-move" together.

The covariance is negative if when X is above it's mean, Y tends to be below it's mean and vice versa. They move in opposite directions, but are still related! In other words, if I know X was above it's mean, then I would predict that Y is below it's mean (knowing one gives me information on the other).

With some algebra and the rules of expectations, we can see

$$\begin{aligned}
\text{cov}(X, Y) &= \mathbb{E}\left[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\right] \\
&= \mathbb{E}\left[XY - \mathbb{E}(X)Y - X\,\mathbb{E}(Y) + \mathbb{E}(X)\,\mathbb{E}(Y)\right] \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y) - \mathbb{E}(X)\,\mathbb{E}(Y) + \mathbb{E}(X)\,\mathbb{E}(Y) \\
&= \mathbb{E}(XY) - \mathbb{E}(X)\,\mathbb{E}(Y).
\end{aligned}$$

The correlation is just a rescaled version of the covariance:

$$\text{corr}(X, Y) = \frac{\text{cov}(X, Y)}{\text{sd}(X)\text{sd}(Y)}.$$

The correlation is designed to always be between -1 and 1 (because of the rescalling), so it is more popular since people are used to thinking about correlations. A correlation close to 1 and/or -1 is a *very strong* relationship.

It is important to know that covariance and correlation only measure a *linear* relationship between X and Y; if the function connecting X and Y is non-linear, then the correlation is a bad summary statistic of the corelationship between them. This is similar to how the mean is a bad measure for highly skewed data.

Independence

Two random variables are said to be **independent** if knowing information about the realization of one of them tells you nothing about the realization of the other one. That is, if I told you the day was sunday ($X = \text{Sunday}$), then you would not have any better prediction about whether it is raining ($Y = 1$). The two random variables are independent.

When X and Y are indepedent, this can be summarized by $f_{X|Y}(x \mid y) = f_X(x)$. This also implies that when X and Y are independent,

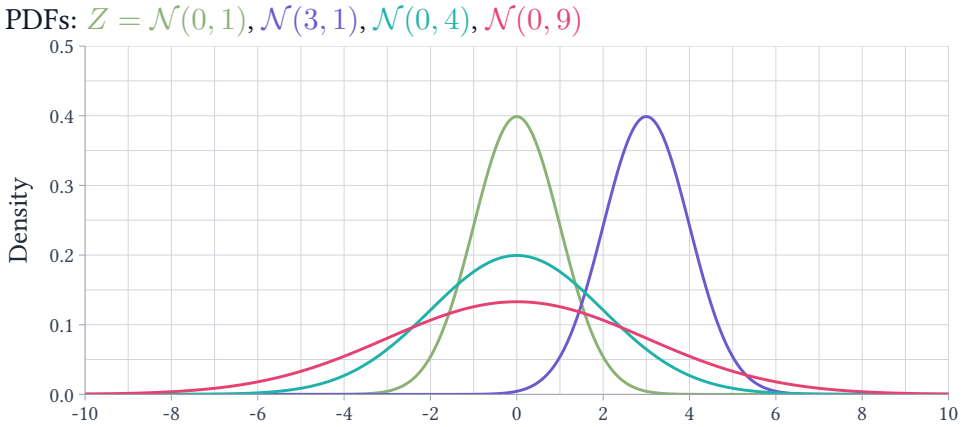
$$F_{X,Y}(x, y) = F_X(x)F_Y(y).$$

Using the definition of expectations, we can derive the following when X and Y are independent:

- $\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y)$
- $\text{cov}(X, Y) = 0$
- $\text{var}\,(aX + bY) = a^2\,\text{var}\,(X) + b^2\,\text{var}\,(Y)$

The last fact comes from the fact that for *all* random variables $\text{var}\,(aX + bY) = a^2\,\text{var}\,(X) + b^2\,\text{var}\,(Y) + 2ab\,\text{cov}(X, Y)$, but the last term is zero from independence.

Figure 1 — Example PDFs for different Normal Distributions



3 — Normal Distribution

The normal distribution is one of the most important distributions in statistics. When a variable is normally distributed, we write $X \sim \mathcal{N}(\mu, \sigma^2)$, where $\mu = \mathbb{E}(X)$ and $\sigma^2 = \text{var}(X)$. The normal distribution is symmetric and the density function of the normal distribution looks like a 'bell'. The parameter μ changes where the center of the distribution is and σ^2 determines how wide the distribution is.

See the figure for example:

We can calculate probabilities of taking values $X \in [\underline{x}, \bar{x}]$ by integrating the area under the pdf. The pdf is given by

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right).$$

Would you want to integrate this function? My guess is probably not.

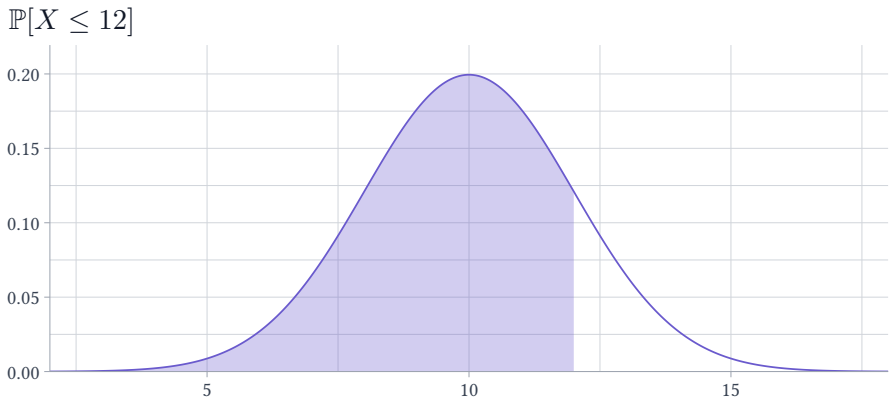
Instead, we refer to the Z-table which calculates probabilities for the **standard normal distribution** denoted $Z = \mathcal{N}(0,1)$. It then calculates $P(Z \leq z)$ for values of z ranging from -3 to 3. Since we do not have a *standard* normal distribution, we can **standardize** our variable X to make it standard normal:

$$\frac{X-\mu}{\sigma} \sim Z = \mathcal{N}(0,1)$$

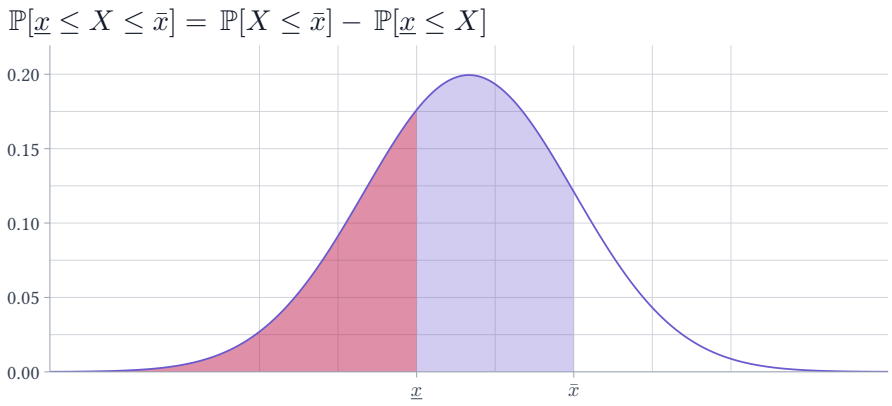
Example 3.1. Say we have $X \sim \mathcal{N}(10, 4)$ and we want to know $\mathbb{P}(X \leq 12)$. Then we can standardize our problem:

$$\mathbb{P}(X \leq 12) = \mathbb{P}\left(\frac{X - 10}{2} \leq \frac{12 - 10}{2}\right) = \mathbb{P}(Z \leq 1)$$

The last probability we can find in our Z-table by looking up the z-score of 1.



Moreover, arbitrary intervals like $[\underline{x}, \bar{x}]$ can be calculated as $\mathbb{P}(X \leq \bar{x}) - \mathbb{P}(X \leq \underline{x})$.

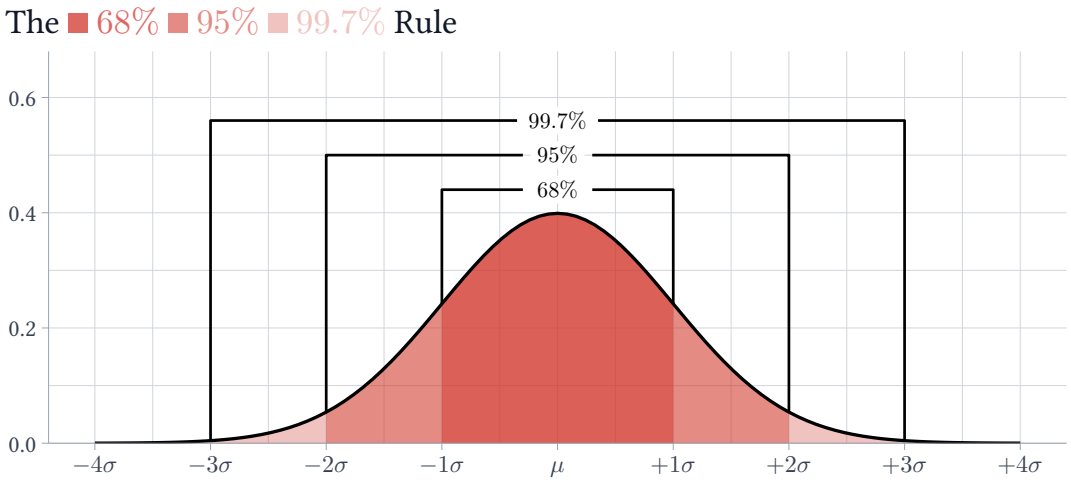


■

The 68-95-99.7 rule

With the normal distribution, we have the 68-95-99.7% rule which says that 68% of observations drawn from a $\mathcal{N}(\mu, \sigma^2)$ random variable will be within the range $[\mu - \sigma, \mu + \sigma]$ (within one

standard deviation from the mean); 95% will fall within two standard deviations from the mean; and 99.7% will fall within three standard deviations. In figure form:



4 — Statistical Inference

Statistical Inference is the procedure of using a random sample of observations from a population to try and learn something about the population distribution of the data. To be more specific, there is some *statistic* of the population distribution (e.g. it's expectation, it's variance, the 80th percentile, etc.) that we want to know about. Let's call that statistic θ.

Estimators

We do not observe the full population; only just a sample. We take our random sample, $X_1, \ldots, X_n$, and want to **estimate** the population statistic. We calculate an **estimate** of the population statistic using the data. The calculation we chose to make is called our **estimator**. For , we can use the sample average from a set of observations to infer about the expectation of the population. To do so, we could calculate the sample average of our sample:

$$\underbrace{\bar{X}}_{\text{Estimator}} \equiv 1/n \sum_{i=1}^{n} X_i$$

Any particular way to estimate the population statistic is called a **estimator**. For any statistic,

there are many different estimators. Therefore, we have some ways to discuss what makes a good estimator.

Property 1: Unbiasedness

We say an estimator W is an unbiased estimator for θ if:

$$\mathbb{E}(W) = \theta.$$

That is, if we take a bunch of random samples (**repeated sampling**) and averaged them, on average the estimator would equal the population statistic.

Note this is not saying the estimate always equals the population statistic, just that on average it does. For , the sample mean does not always equal the population mean, but it is an unbiased estimate for the population mean.

Property 2: Consistency

Let W_n be an estimator of θ based on a population of size n. The estimator is consistent if as n gets larger, $W_n \to \theta$ and the variance of W_n shrinks to 0. That is, the estimator gets more precise and centers around the population statistic.

Property 3: Efficiency

If W_1 and W_2 are two unbiased estimators, we say W_1 is more efficient than W_2 if $\text{var}\,(W_1) < \text{var}\,(W_2)$.

Confidence Intervals

Since the estimator does not always equal it's population statistic (in repeated sampling) We want to be able to describe the uncertainty of an estimator. To do so, we need to have an estimate to the variability of our estimator in repeated sampling.

That is, we want to quantify how much we think our estimator would move around if we were to collect a bunch of different samples of the same size and calculate the estimator for each sample. This thought experiment of grabbing a bunch of samples of the same size and calculating our estimator on each sample creates what we call our **sample distribution** of our estimates.

It turns out that for a lot of estimators, the sample distribution is approximately **normally distributed**. That makes summarize our uncertainty around our estimator easier. We typically

will construct a **confidence interval** to descirbe our level of certainty. The confidence interval is centered on our estimate, but it's width describes a range of values we think the population statistic could be.

Example 4.1. Let $\{Y_1, \ldots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \text{var}(Y)$.

We can construct a 95% confidence interval as:

$$\mathbb{P}\left(-1.96 < \frac{\bar{Y} - \mu}{\sigma/\sqrt{(n)}} < 1.96\right) = 0.95$$
$$\implies \mathbb{P}\left(\bar{Y} - 1.96 * \frac{\sigma}{\sqrt{(n)}} < \mu < \bar{Y} + 1.96\frac{\sigma}{\sqrt{(n)}}\right) = 0.95$$

That is, in repeated sampling, there is a 95% probability (95 out of 100 samples of size n) that μ falls within our confidence interval:

$$\left[\bar{Y} - 1.96 * \frac{\sigma}{\sqrt{(n)}}, \bar{Y} + 1.96 * \frac{\sigma}{\sqrt{(n)}}\right]$$

Since we do not observe σ, we will estimate it using the following estimator, called the sample standard deviation of y:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

In the above procedure, the values -1.96 and 1.96 are the critical values for the 2.5th percentile and 97.5th percentile of the standard normal distribution, i.e. we have the middle 95%. If we want to be more confident in our interval, then we need to use larger critical values. For , if we want to be 99% confident in our interval (have θ in our interval in 99 out of 100 samples), then we need to use the 0.5th percentile and 99.5th percentile of the normal distribution for our critical values.

Note this procedure works because $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$ and therefore our Z-**score**, $\frac{\bar{Y}-\mu}{\sigma/\sqrt{(n)}}$, is approximately distributed as the standard normal distribution.

Hypothesis Testing

In addition to confidence intervals, it is common to perform **hypothesis tests**. Simply put, we want to test whether evidence is consistent with a **null hypothesis** that θ equals some value (e.g.

we could hypothesize that the population mean height is 5 ft. 8 in.). We then would calculate our sample mean and if it is "too far" from the null hypothesis population mean, then we say the evidence rejects the null hypothesis.

How far is "too far" away from the null hypothesis? That depends on (1) how noisy our estimate is (the sample distribution) and (2) how confident we want to be in rejecting the null.

When I said before that we will see if our sample mean is "too far" away from the null, that was a little bit incorrect. For reasons we will not cover, we will transform the sample mean into a **test statistic**. When we compute the statistic for a particular outcome, we obtain an outcome of the test statistic, which we will denote T.

When the null hypothesis is true, the test statistic will be distributed as $\mathcal{N}(0, v)$ (normally distributed with mean 0 and some variance v). Therefore, we expect T to be very close to zero (if the null is true). So, if we find a value of $T = t$ that is "very far" away from zero, then we find evidence against the null hypothesis.

Given a level of confidence (e.g. 95%) and the variance v, we can use properties of the normal distribution to determine the probability that a draw from the $\mathcal{N}(0, v)$ distribution would be $\geq \left|\hat{t}\right|$. This is called the p-**value**. In words, the p-value tells us, assuming the null hypothesis is true, how often would we expect to see a value as large or larger than the one we *did observe in our sample*. If that probability is small, then we will reject the null.

In particular, given our level of confidence (e.g. 95%), define the **significance level** as $\alpha = 1-$ the level of confidence (e.g. 5%). We reject the null if the p-value, p is smaller than the level of confidence. Typically, we will use $\alpha = 0.05$.

Example 4.2. Let $\{Y_1, \ldots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \text{var}\,(Y)$.

We will test the null hypothesis $H_0 : \mu = \mu_0$ for some proposed value μ_0.

Then, our test statistic is given by

$$T = \frac{\bar{X} - \mu_0}{\hat{\sigma}/\sqrt{(n)}},$$

where $\hat{\sigma}$ is the square root of our estimate of the variance of Y. Assuming the null hypothesis is true, H_0, then $T \sim \mathcal{N}(0, 1)$.

Therefore, the p-value is given by

$$p = \mathbb{P}(T \geq |t|) = 2 * \mathbb{P}(Z \leq -|t|),$$

where Z is the standard normal random variable.

You can find this probability by looking up $-|t|$ in the Z-table.

■

Last, we will discuss the **rejection region** for a given *null hypothesis*. This is defined as the set of values of our estimate where we would reject the null hypothesis for a given significance level, α.

Example 4.3. Let $\{Y_1, \ldots, Y_n\}$ be a random sample. The sample mean is approximately distributed $\mathcal{N}(\mu, \sigma^2/n)$ where $\mu = \mathbb{E}(Y)$ and $\sigma^2 = \text{var}(Y)$.

We will test the null hypothesis $H_0 : \mu = \mu_0$ for some proposed value μ_0.

The rejection region can be found, most simply, by forming a confidence interval around the null hypothesis. Any value *outside this interval* will be rejected.

Our confidence interval for $\alpha = 0.05$ is given by

$$\left[\mu_0 - 1.96 * \sigma/\sqrt{n}, \mu_0 + 1.96 * \sigma/\sqrt{n}\right]$$

Therefore our rejection region is $\bar{X} \leq \mu_0 - 1.96 * \sigma/\sqrt{n}$ or $\bar{X} \geq \mu_0 + 1.96 * \sigma/\sqrt{n}$.

■

Review

In the population of cities granted enterprise zones in a particular state, let Y denote the percentage change in investment from the year before to the year after a city became an enterprise zone. Assume that Y has a $\mathcal{N}(\mu, \sigma^2)$ distribution.

- State the null hypothesis that enterprise zones have no effect on business investment θ State the alternative hypothesis that they have a positive effect
- Suppose that the sample yields $\bar{Y} = 8.2, s = 23.9$ and $n = 36$. Construct and calculate the test statistic. Do you reject the null?

- What is the rejection region for a 5% significance level?