Probability and z-scores

Probability Basics

$$P(A) = \frac{\text{number of times } A \text{ occurs}}{\text{number of possible outcomes}}$$

Probability Rules:

$$P(A) = 1 - P(\text{not } A)$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

If the events are independent...

$$P(A \text{ and } B) = P(A)P(B)$$

In general...

$$P(A \text{ and } B) = P(A|B)P(B)$$

Where $p(A|B)$ is the probability of event A given that event B occurred. It's important to remember that 1) probabilities always sum to 1, and 2) probability can also be thought of as an area under a distribution.

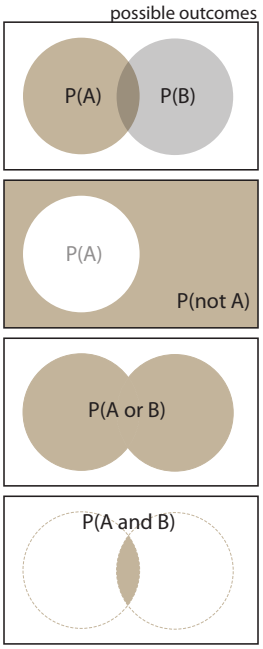

Figure 1: Visual probability rules.

z-scores

z-scores are a way of standardizing from different distributions.

Population z-score: $$z = \frac{X - \mu}{\sigma}$$

Sample z-score: $$z = \frac{X - \bar{X}}{s}$$

If X follows a normal distribution then the z-scores follow a *standard normal distribution* (the normal distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$), and probabilities for any interval of the distribution can be looked up using the standard normal table.

z	$p(-z \leq x \leq z)$
1	0.683
2	0.955
3	0.997

z	$p(x \leq z)$
1.645	0.950
2.326	0.990
3.901	0.999

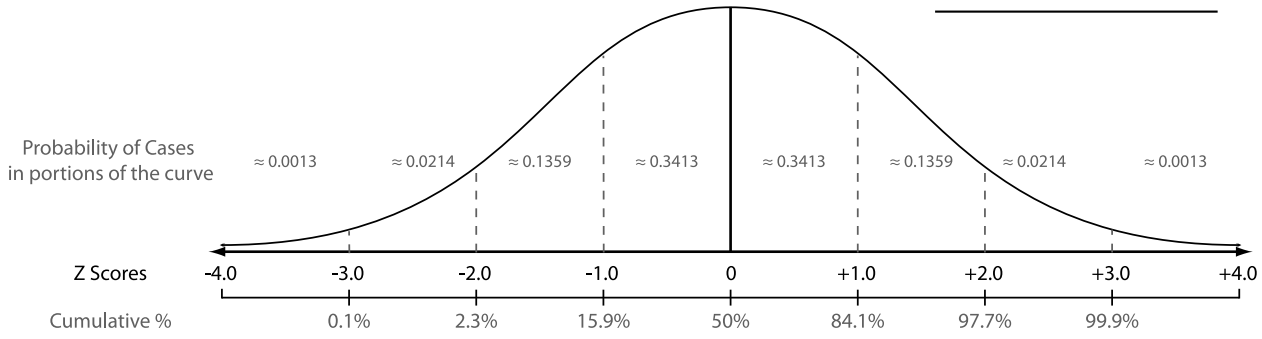

Figure 2: Standard Normal Distribution with mean $\mu = 0$ and standard deviation $\sigma = 1$. z-scores measure deviations from the mean in units of standard deviations.

Distribution of Sample Means

In many cases, instead of comparing single scores, we want to summarize the entire sample and compare the average outcome to some known quantity. Since samples vary, we need a way to think about how certain we are about the sample mean. What would we expect if we ran the same experiment again? The answer is that we would get the distribution of sample means, which has two key properties

1. The mean and standard deviation are related to the population parameters with $\mu_M = \mu$ and $\sigma_M = \sigma/\sqrt{n}$.
2. Surprisingly, no matter what the shape of the population is, the distribution of sample means becomes normal for $n > 30$.

The purpose of this distribution is to describe the expected variation in the mean due to chance sampling error.

The z test (one-sample location test)

Describing *what happens by chance* is the key to all of hypothesis testing. Suppose, we want to make a decision about whether the mean of our data is significantly different from some known value. Assuming that the population is normal or $n > 30$ and that σ is known, the distribution of sample means exactly describes what we expect. After calculating a test statistic $z = \frac{\bar{X}-\mu}{\sigma_M}$ we can look up a probability (p-value) for how extreme such a result would be in a standard normal table.

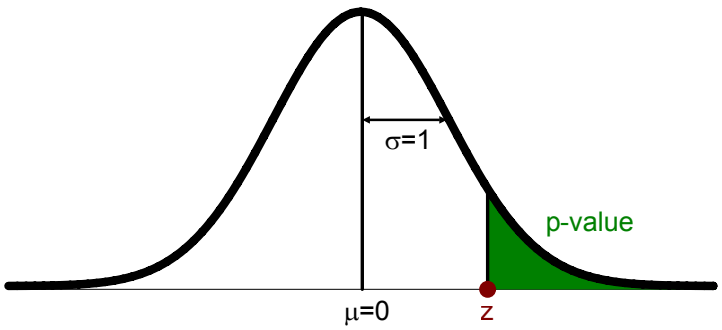

Figure 3: Illustration of a one-tailed z-test. The p-value (shaded) is the probability of getting a result that is as extreme or more extreme than the test statistic (assuming the null hypothesis is true).

In almost all real world situations we don't know the parameters of the population. Instead, we estimate σ_M using the *standard error of the mean*:

$$SEM = \frac{s}{\sqrt{n}}$$

Errorbars are commonly reported/plotted using $\bar{X} \pm SEM$.