Random Sampling: Ex. Process of drawing numbers at random

Cumulative Distribution Function (cdf) The probablity that a random variable X takes on a value less than or equal to x

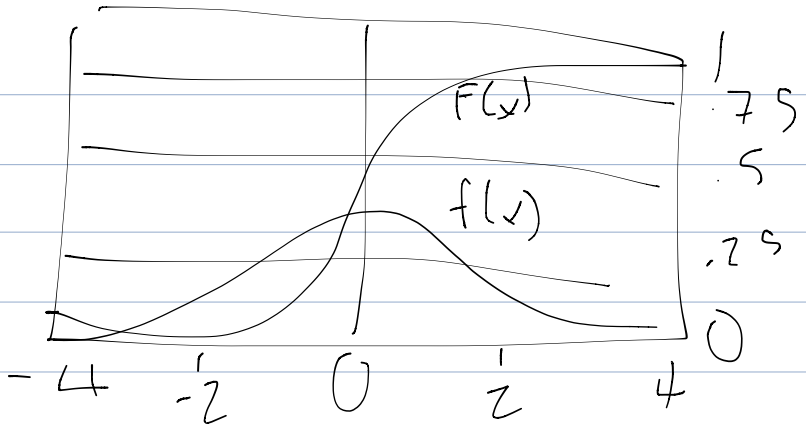
We denote this as F(x) = P(X <= x)

Example: If the heights of 40% of a population are less than or equal to 70 inches, then the probability that a randomly selected height X is less than or equal to 70 is .4

F(70) = .4

Probability Density Function (pdf): If we are measuring continuous variables, then probabilities can be expressed as areas under a curve f(x) called the probability density function

- Ex. of continuous variables



Normal Distribution

Mu:

The mean of the distribution (mean of X)

sigma:

The standard deviation of the distribution (standard deviation of X)

Standard Normal Distribution

Normal distribution where mu=0 and sigma=1

N(0,2) = mu=0, variance/sigma^2=2

Normal Distribution

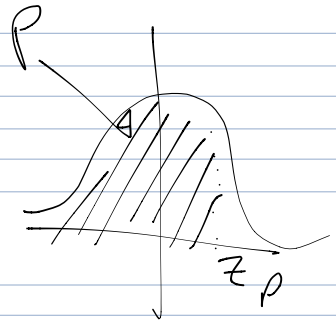
-z_p:

The (100p)th percentile of the standard normal distribution

-P(Z <= z_p) = p

-Common Values for z_p:

z_.90 = 1.282, z_.95 = 1.645, z_.975 = 1.96



Normal Distribution

-z_p:

Alternatively, we can index these percentiiles by their upper-tail probabilities as z(a), where a is the upper-tail probability

Ex.: z(.05) = z_.95 = 1.645

X_bar:

The sample mean

What is the difference between X_bar and x_bar?

X_bar = random variable of the sample mean

x_bar = value of the sample mean

Sample of 100 people from large population
x_bar1, x_bar2, ..., x_barN

Central Limit Theorem

If a simple random sample of size n is drawn from any population with mean mu and finite standard deviation sigma, then when n i sufficiently large enough, the sampling distribution fo the sample mean is approximately Normal with mean mu and standard deviation sigma/sqrt(n): N(mu, sigma^2/n)

What is notable if the simple random sample is drawn from a population with a Normal distribution?

A type of family of distributions

f(x) = 1/b * h * (x-a / b)

where h(z) is a standard form of the distribution, a is the location parameter, and b is the scale parameter

The location parameter has the effect of shifting the distribution, and the scale parameter scales the distribution

Ex. The normal distribution: a = mu, b = sigma, and h(z) is the standard normal distribution

Tails

- Tail Weight:

Heavy-tailed distribution: Will occasionally prouce observations that are much more extreme than others (Laplace, exponentional, and Cauchy)

Note: The Cauchy distribution has such heavy tails that the man and variance of the distribution do not exist

Light-tailed distribution: Do not produce these relatively extreme observations (uniform and normal)

- Skewness:

A measure of the asymmetry of the distribution

The uniform, normal, Laplace, and Cauchy distributions are symmetric. The exponential distribution is skewed to the right.

Binomial Setting

- There are a fixed number of n observations
- The n observations are independent
- Each observation can be classified as one of two categories, a "success" or a "failure". The usual way of noting this is with a 1 for a "success" and a 0 for a "failure"
- Each observation has a constant probability of success, p. p is the same for each observation

- A state's has a population of voters, and 40% like a specific candidate for elected office, while 60% do not. In a sample of those voters, we may associate a 1 with those voters who like the candidate and a 0 for those who do not.
- What is p here? 0.4
- Are all of the setting assumptions met? Ask for fixed number, then yes

- Suppose we select n elements randomly from a population, and we let X denote the number of "successes" (i.e. The number of 1's). If the other binomial setting assumptions hold, then the probability distribution of X is called the binomial distribution and has the following form:

f(x) = (n x) * p^x * (1-p)^(n-x), x = 0,1,...,n

What is the mean of X? np
What is the variance of X?

- Note: For large n, the binomial distribution can be approximated by the normal distribution with the same mean and variance

$$Z = \frac{X - \mu}{\sigma} \quad \mu \leftarrow np, \quad \sigma \leftarrow \sqrt{npq} \leftarrow (1-p)$$

Confidence Interval Review:

- Suppose we wish to estimate the mean of a population. The sample mean based on a random sample of size n called a point estimate of the population mean.
- Let us also suppose, that the population from which a sample is to be taken is normally distributed with a variance of sigma^2
- Is mu known?
- Remember, Z is a standard normal random variable:

  Z = X_bar - mu / sigma/sqrt(n)

- Since Z has the property P(-1.96 < Z < 1.96) = .95, then:

$$P = \left(-1.96 < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < 1.96\right) = .95$$

- We can obtain a 95% confidence interval by solving for mu:

$$\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}$$

- More generally we can obtain a (100(1-a))% confidence interval in this setting by:

$$\bar{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$$

- What is a 95% Confidence Interval?

  95% of the Confidence Intervals created will contain the true mean

- Example: Suppose it is known that the heights of the male population have a normal distribution with a standard deviation of 3. Suppose in a random sample of size 20 from the population we find that the sample mean is 70.8 inches. What is the 95% confidence interval for the mean height of this population?

$$70.8 \pm 1.96\frac{3}{\sqrt{20}}$$

$$70.8 \pm 1.3148$$

$$[69.485, 72.115]$$

Tests of Hypotheses Review

- A statistical test of hypothesis is a procedure for deciding between two hypotheses called the null hypothesis H_0 and the alternative hypothesis denoted H_a.
- An upper-tail test of hypothesis about the mean of a population has the form:

H_0 : mu = mu_0, H_a : mu > mu_0

- What would be the lower-tail form for a similar test? mu < mu_0
- What are they called upper and lower tail tests? Going to be the same

- For an upper-tail test of hypothesis about hte mean of a population where the standard deviation si known, one would reject the H_0 in favor of H_a at a a level of significance if:

- Usually, one begins analysis by determining whether or not it is appropriate to use normal-theory methods. In other words, one begins by deciding whether or not the normal distribution is an appropriate model for describing the population of interest.
- If not initially, sometimes transforming data may yield data that is more suitably described by the normal distribution. Then normal-theory methods may be applied.
  Ex.: Y = log(X) may suitably transform non-normal X data into normally distibuted Y data.

- If normal-theory methods are not appropriate to either the original or transformed data, one may try to identify the underlying, non-normal, distribution of the data, and then use

Nonparametric methods require minimal assumptions about the form of the distribution of the population of interest.

Ex.: Assume a continuous distribution, but no other assumptions
Ex.: Assume the population distribution depends on location and scale parameters, but no other information

On the other hand, parametric methods require that the form of the population distribution be completely specified except for a finite number of parameters
Ex.: A one-sample t-test for means assumes sampling.....

11/13/16 Tues