

Probability Distributions and the Normal Distribution

A **Random Variable** is a function or rule that assigns a number to each outcome of an experiment. This could be the number of heads observed in an experiment that flips a coin 10 times, or the annual revenue of a randomly selected hospital. Random variables can be discrete or continuous.

A **Probability Distribution** is a listing of all possible numerical outcomes from a random variable. There are discrete and continuous probability distributions. The binomial distribution and Poisson distribution are examples of discrete probability distributions. Chapter 5 in the text deals with these. We are skipping for sake of time.

Continuous Distributions

When we have a continuous RV we have another set of distributions to draw from. Since there are an infinite number of possible outcomes, one characteristic of a continuous distribution is that the probability of one exact value is zero, (probability of being exactly 6 feet tall is zero) thus we tend to measure things in intervals (between 5'9" and 6'0").

The Normal Distribution

One particularly useful distribution is the normal distribution. We will to a lot with this distribution throughout the remainder of the semester, thus the book devotes an entire chapter to it.

Properties of the normal distribution:

1. It is bell shaped (and thus symmetrical) in appearance
2. its measures of central tendency are all identical
3. Approximately 95% of all occurrences of the random variable occur within two standard deviations of the mean.
4. Its associated random variable has an infinite range.

Draw some normal distributions

We use the expression $f(X)$ to denote the probability density function. For the normal distribution the density function is:

$$f(X) = \frac{1}{\sqrt{2\pi}\sigma} * e^{-(1/2)[(X-\mu)/\sigma]^2}$$

So if we want to know the probability that a normally distributed random variable is between $-\infty$ and X we would just plug X into the above equation and it would tell us.

Note that since μ and σ will change for each possible distribution there are an infinite number of these. So we would have to make this calculation every time we want to know something. One way around this is to transform the variable X to a specific normal distribution.

Suppose we Let:

$$Z = (X - \mu) / \sigma$$

Now this new variable Z will be normally distributed with a mean of zero and a standard deviation of 1.

Note that what this is doing is putting the variable in standard deviation units.

So if we have a normal distribution with a mean of 5 and a SD of 2 and we are concerned with the outcome 3, then

$$Z = (3 - 5) / 2 = -1. \text{ In other words 3 is one standard deviation below the mean.}$$

Or if we are interested in the outcome 9:

$$Z = (9 - 5) / 2 = 2 \text{ or 9 is two standard deviations above the mean.}$$

Now the density function for the standard normal distribution simplifies to:

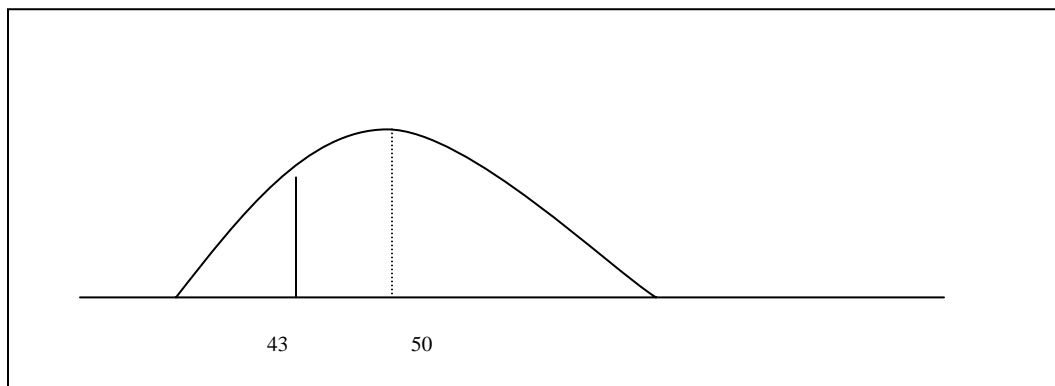
$$f(Z) = (1/\sqrt{2\pi})e^{-(1/2)Z^2}$$

This is easier to deal with since everything but Z is a constant.

EX:

Given a normal distribution with a mean of 50 and a sd of 4, what is the probability that:

1. $X < 43$



First convert to Z:

$$Z = 43 - 50/4 = -7/4 = -1.75$$

We can either look in a standard normal table or use NORMSDIST(Z) in Excel

We get .0401, or there is about a 4 percent chance that x is less than 43.

2. what is the probability that $x > 43$?

$$1 - .0401 = .9599$$

3. $42 < X < 48$

First draw the picture:

Calculate Zs

$$Z_{42} = 42 - 50/4 = -2$$

$$Z_{48} = 48 - 50/4 = -.5$$

$$P(X < 42) = .0228$$

$$P(X < 48) = .3085$$

$$\text{So } P(42 < X < 48) = .3085 - .0228 = .2858$$

4. $X > 57.5$?

$$Z = 57.5 - 50/4 = 1.875$$

$$P(X < 57.5) = .9696$$

$$\text{So } P(X > 57.5) = 1 - .9696 = .0304$$

Sampling Distributions

One of the main goals of data analysis is to make statistical inferences – that is to use information about a sample to learn something about the population.

Census vs CPS

To make life easier we often take a sample and use the sample statistics to say something about the population statistics, but in order to do this we need to say something about the sampling distribution. That is, there are lots of potential samples we could take of a population and each would give us slightly different information. Dealing with this sampling distribution helps us to account for this.

The Sampling Distribution of the Mean

Now we want to talk about making inferences about the mean of a population

The sample mean is an **unbiased** estimator of the population mean – this means that if we took all possible sample means and averaged them together, the average would equal the population mean.

Population		
Student	GPA	
A	3.82	
B	3.55	
C	2.89	
Mu	3.42	
Sigma	0.3906405	
	Sample	Xbar
AA		3.82
AB		3.685
AC		3.355
BA		3.685
BB		3.55
BC		3.22
CA		3.355
CB		3.22
CC		2.89
	Mux-	3.42
	Sigma(xbar)	0.276225
	Sigma/Sqrt(n)	0.276225

So $(6.5-6.25)/(3.62/5) = .345$, converting this to a probability ($\text{normdist}(.345)=.635$, $1-.635=.365$). This tells us that there is a 36.5% chance of getting a sample mean of 6.5 hours or more.

Note that since the square root of n is in the denominator of the standard error, as the sample size increases the standard error will decrease, making any given sample mean less likely.

If $n=100$

$Z = (6.5-6.25)/3.65/10 = .691$, or only a 24.5% chance.

Sampling from Non-Normally Distributed Populations

What we've done above assumed that we were sampling from a normally distributed population. But often the population distribution is not normal or is unknown. The good news is the central limit theorem:

Central Limit Theorem -- as the sample size gets large, the sampling distribution of the mean can be approximated by the normal distribution.

As a general rule of thumb sample sizes of at least 30 will tend to give a sampling distribution that is approximately normal.

The sampling distribution of the proportion

Sometimes the random variable we are interested in is based on a categorical response – yes/no, male/female, etc. What is done here is to assign one group 1 and the other 0 then the sample mean would be the sum of the ones, which gives the proportion of the sample in that category

The sample proportion $P_s = X/n$
 X is the number of successes, n is the sample size

Again the sample proportion is an unbiased estimator of the sample proportion and the standard error of the proportion is given by:

$$\sigma_{ps} = \sqrt{(p(1-p)/n)}$$

The sampling distribution of the proportion follows the binomial distribution, but the good thing here is that we can use the CLT and use the normal distribution

So that

$$Z = \frac{P_s - P}{\sqrt{(p(1-p)/n)}}$$

Historically 10% of a large shipment of machine parts are defective. If random samples of 400 are selected, what proportion of the samples will have: less than 8% defective?

$$Z = \frac{.08 - .10}{\sqrt{(.1(.9)/400)}}$$

$= -.02 / .015 = -1.33 \sim$ that about 9.1 % of samples will have less than 8% defective

$$24 - (1.96)(30/\sqrt{100}) < \mu < 24 + (1.96)(30/\sqrt{100})$$

$$1.96(30/10) = 5.88$$

$$\text{so } 18.12 < \mu < 29.88$$

So we are 95% confident that the population mean wage is between 18.12 and 29.88.

What if we want a 90% confidence interval?

Draw this

Need to find z value corresponding to .05 = use = Normsinv(.05) = 1.645

$$24 - (1.645)(30/\sqrt{100}) < \mu < 24 + (1.645)(30/\sqrt{100})$$

$$19.065 < \mu < 28.935$$

A tighter interval. Note though that the cost of the tighter interval is a lower level of confidence.

Confidence Interval Estimation of the Mean when σ is unknown.

The above examples assumed, maybe unrealistically, that the population standard deviation was known. In most cases if we do not know the population mean we will not know the standard deviation either. How do we handle this.

If a random variable X is normally distributed, then the following statistic has a t distribution with $n-1$ **degrees of freedom**:

$$t = (X - \mu) / (S / \sqrt{n})$$

The t -distribution looks much like the normal distribution except that it has more areas in the tails and less in the center. Draw this.

Because σ is unknown and we are using S to estimate it, then we are more cautious in our inferences, the t -distribution takes this into account.

As the sample size increases, so do the degrees of freedom, and so the t -distribution becomes closer and closer to the normal distribution

Degrees of Freedom

Recall that in calculating the sample variance we did:

$$\sum (X_i - \bar{X})^2$$

note that in order to estimate S^2 we first had to estimate \bar{X} , so only $n-1$ of the sample values are free to vary.

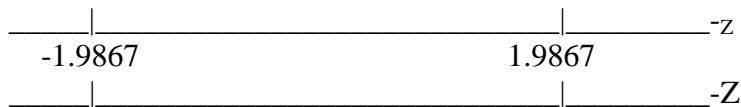
Suppose a sample of five values has a mean of 20. we know $n=5$ and $\bar{X}=20$ so we know that $\sum X_i=20$

Thus once we know four of the five values, the fifth one will not be free to vary because the sum must add to 100. if four of the values are 18, 24, 19, and 16 the fifth value has to be 23.

Estimating confidence intervals with the t-distribution

In a study of costs, it was found that the hospital costs of managing 90 cases of heart disease averaged 1120 with a standard deviation of 75. Construct a 95% confidence interval for the true average cost of managing this condition.

First draw the picture



Excell Commands:

`tdist(x,df,tails)`

`tinv(prob,df)` -- assumes two tails

$$\text{so } 1120 - (-1.9867) * (75/\sqrt{90}) < \mu < 1120 + (-1.9867) * (75/\sqrt{90})$$

$$\text{or } 1,104.5 < \mu < 1,135.5$$

Suppose we know the average length of stay experience by 100 patients was 3.5 days and the standard deviation was 2.5 days. Construct a 98% CI for the true mean stay.

$$Z = 2.3642 \text{ so}$$

$$3.5 - (2.3642) * (2.5/\sqrt{100}) < \mu < 3.5 + (2.3642) * (2.5/\sqrt{100})$$

$$\text{or } 2.91 < \mu < 4.09$$

Confidence interval estimation for the proportion

Here we can do exactly the same thing

$$P_s - Z\sqrt{(P_s(1-P_s)/n)} < P < P_s + Z\sqrt{(P_s(1-P_s)/n)}$$

Suppose we are interested in the proportion of the RN population who are members of a union. A sample of 200 RNs is taken and 5% are found to be unionized. Construct the 95% CI for the true proportion of unionized RNs.

$$.05 - 1.96\sqrt{(.05)(.95)/200}$$

$$.05 \pm 1.96(.0154)$$

or we are 95% confident that the population proportion of RNs who are unionized is between .020 and .080

Determining the Sample Size

Up to now the sample size has been predetermined, but in reality the researcher can often set the desired sample size. Note that the larger the sample size the lower your sampling error, but also the higher the cost of obtaining the information.

$$\text{Recall: } Z = (X - \mu) / (\sigma / \sqrt{n})$$

$$\text{Or } X - \mu = Z(\sigma / \sqrt{n})$$

Or this second term is known as the error – how far off the sample mean is from the pop mean.

$$e = Z(\sigma / \sqrt{n})$$

We can solve this for n to get:

$$n = Z^2 \sigma^2 / e^2$$

so you can pick your sample size to get the desired level of error. It is a function of:

1. the desired level of confidence, the Z critical value
2. the acceptable sampling error, e
3. the standard deviation

Suppose we want to estimate the average salary paid to RNs and we want our estimate to be off by less than \$100 with a probability of .95. Suppose we think the population standard deviation $\sigma = \$1,000$. How big of a sample do we need?

$$n = 1.96^2 (1,000)^2 / 100^2 = 384.2. \text{ So a sample of 384 will work.}$$