

# Stat 450

## Chapter 7: Sampling distributions

Fall 2016

We turn now in full force to studying **sampling distributions**. Let's define some terms:

- **Population:** Collection of all elements of interest.
- **Parameter:** A quantity (or quantities) that, for a given population, is fixed and that is used as the value of a variable in some general distribution or frequency function to make it descriptive of that population. (E.g.,  $\mu$  is the mean of a normal distribution;  $(\alpha, \beta)$  govern the Gamma distribution and are used to define the mean and variance.)
- **Sample:** A collection of elements drawn from the population and observed. In these notes, we will be considering univariate realizations  $Y_1, Y_2, \dots, Y_n$  drawn independently and identically from the population.
- **Statistic:** A function of the observable random variables in a sample.

These are important, because they bring us to the definition of a **sampling distribution**. A **sampling distribution** is the distribution of a statistic across repeated samples taken from the population. What are sampling distributions used for?

- Finding the mean and variance of a statistic; this is how *bias* and *mean-squared error* are defined (more later)
- Hypothesis testing
- Finding confidence intervals

### Facts when sample is drawn from any distribution:

1.  $\bar{Y}$  is independent of the residuals  $(Y_1 - \bar{Y}), (Y_2 - \bar{Y}), \dots, (Y_n - \bar{Y})$ . This fact is best proved with linear algebra (see Chapter 13). However, any given residual is not independent of another residual; i.e.  $(Y_i - \bar{Y})$  is not independent of  $(Y_j - \bar{Y})$ , since  $\sum_{i=1}^n (Y_i - \bar{Y}) = 0$ .
2. The previous point implies that  $\bar{Y}$  and  $s^2$ , the sample variance, are independent.
3.  $E(\bar{Y}) = \mu$  and  $Var(\bar{Y}) = \sigma^2/n$ . Here,  $\mu$  is the mean of the distribution and  $\sigma^2$  is the variance of the distribution (note these might be functions of other parameters governing the distribution; e.g.  $\mu = \alpha\beta$  if the distribution is  $GAM(\alpha, \beta)$ ). This is a result from Chapter 4.
4. If  $n$  is "large", then  $\bar{Y} \sim N(\mu, \sigma^2/n)$ . This is a result of the **central limit theorem**. We will prove this at the end of this handout.

### Facts when sample is drawn from a $N(\mu, \sigma^2)$ distribution:

1. **No matter what size  $n$  is**,  $\bar{Y} \sim N(\mu, \sigma^2/n)$ , where  $\mu$  is the mean of the normal distribution and  $\sigma^2$  is the variance of the normal distribution.
2. The following, useful for obtaining confidence intervals and doing hypothesis tests for  $\sigma^2$  :

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$$

3. The following, useful for obtaining confidence intervals and doing hypothesis tests for  $\mu$ :

$$\frac{\bar{Y} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

4. If  $Y_1, Y_2, \dots, Y_n$  is an i.i.d. sample from a  $N(\mu_Y, \sigma_Y^2)$  distribution; and  $X_1, X_2, \dots, X_m$  is a sample drawn i.i.d from a  $N(\mu_X, \sigma_X^2)$  distribution; then the ratio of the sample variances scaled as follows,  $\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$ , follows an  $F_{m-1, n-1}$  distribution.

We will prove facts 1-4 when the sample (or samples, in the case of #4) is drawn from a normal population. We will then prove the central limit theorem, which guarantees normality of  $\bar{Y}$  for large  $n$  when the sample comes from *any* distribution.

**Proof of 1**

## Proof of 2

Preliminaries:

- A. If  $Y \sim N(\mu, \sigma^2)$ , then  $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$  (Will prove.)
- B. If  $Z \sim N(0, 1)$ , then  $U = Z^2 \sim \chi_1^2 \equiv GAM\left(\alpha = \frac{1}{2}, \beta = 2\right)$  (Proved using pdf method in Chapter 6).
- C. If  $Z_1, Z_2, \dots, Z_n$  are i.i.d.  $\sim \chi_1^2$ , then  $\sum_{i=1}^n Z_i \sim \chi_n^2$ . (Will prove.)

(Proof of 2, continued)

Using #2 to derive confidence intervals for  $\sigma^2$

Below is some R code to simulate a sample of size  $n$  from a  $N(0, \sigma^2 = 4)$  population. The function calculates and returns a single 95% confidence interval. We then replicate this function many times to obtain many confidence intervals, 95% of which should cover the true  $\sigma^2$ :

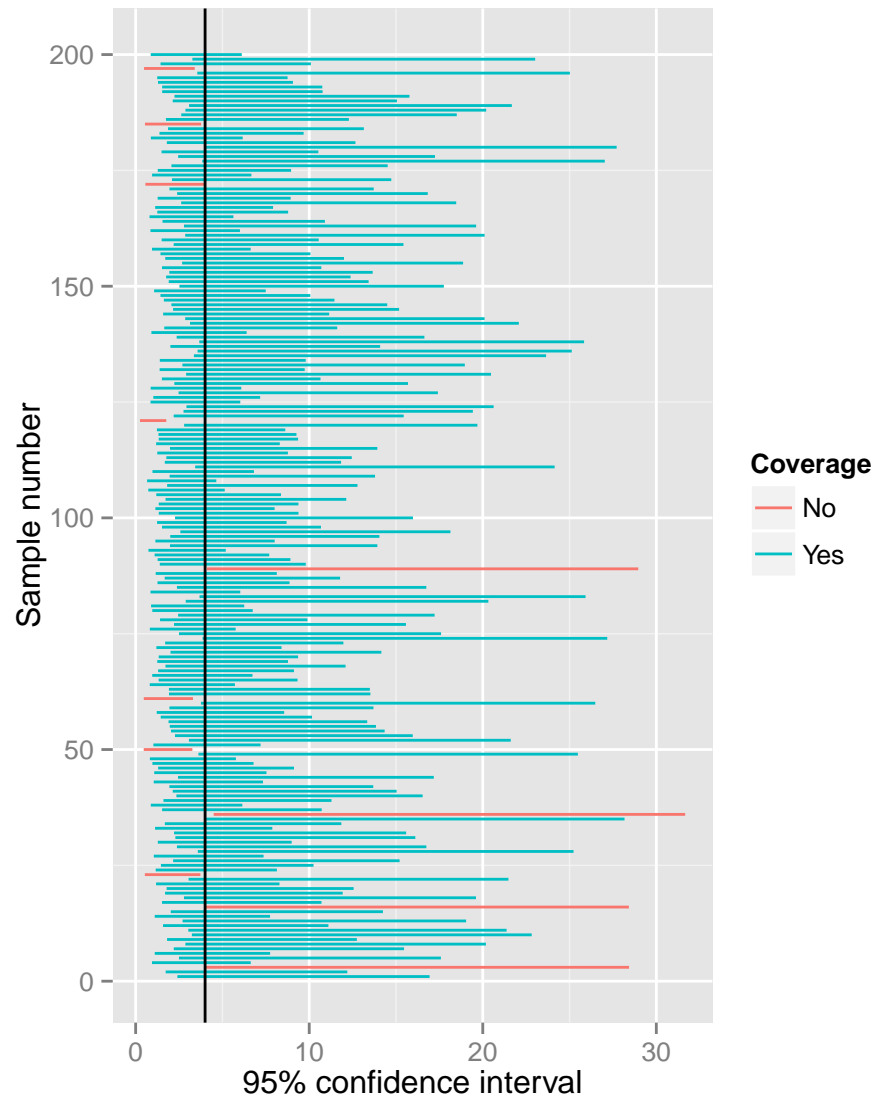
```
#Write code to get sample, calculate 95% confidence interval
get.one.ci <- function(n){
  one.sample <- rnorm(n, mean = 0, sd = sqrt(4))
  s2 <- var(one.sample)
  lower <- (n-1)*s2/qchisq(0.975,n-1)
  upper <- (n-1)*s2/qchisq(0.025,n-1)
  ci <- c(lower,upper)
  return(ci)
}

#Given a 95% confidence interval, and a value of the true sigma^2, does the interval cover sigma^2?
covers.sigma2 <- function(ci,sigma2) {
  cover <- ifelse(ci[1] < sigma2 & sigma2 < ci[2], 'Yes', 'No')
  return(cover)
}

#Gather 200 samples and corresponding confidence intervals, and calculate coverage:
set.seed(24211)
many.ci <- replicate(200, get.one.ci(n=10), simplify='matrix')
df <- data.frame(t(many.ci))
df$Coverage <- apply(df, 1, covers.sigma2, sigma2=4)
df$Sample <- 1:nrow(df)
table(df$Coverage)/200
```

```
##
##      No    Yes
## 0.055 0.945
```

```
##Plot the results
library(ggplot2)
ggplot(data = df) +
  geom_segment(aes(x = X1, xend = X2, y = Sample, yend = Sample,color=Coverage)) +
  geom_vline(xintercept=4) + xlab('95% confidence interval') + ylab('Sample number')
```



### Using #2 for hypothesis testing

EXAMPLE: Quality control. On a production line, consistency of performance is very important. For example, suppose a machine is calibrated to fill 12-ounce Coke bottles very precisely. The machine is supposed to fill each bottle to be 12 ounces, but may have slight variations from bottle-to-bottle. Specifically, suppose the distribution of *actual* bottle fills is intended to follow a normal distribution with mean  $\mu = 12$  and standard deviation of  $\sigma^2 = 0.01$ . If there is evidence that  $\sigma^2 > 0.01$ , the machine will need to be recalibrated. This then becomes a problem of testing:

$$H_0 : \sigma^2 = 0.01$$

$$H_a : \sigma^2 > 0.01$$

Suppose a sample of  $n = 20$  bottles is taken from the production line; how large will  $s^2$  need to convincingly suggest the machine needs to be recalibrated? This involves finding the sampling distribution of  $s^2$  (or some appropriate scaled version thereof), to find what values of  $s^2$  would be very unusual if  $H_0$  were true.



**Proof of #3**

Here, we want to prove that, if  $Z \sim N(0, 1)$ , and  $W \sim \chi_\nu^2$ , that:

$$T = \frac{Z}{\sqrt{W/\nu}} \equiv \frac{N(0, 1)}{\sqrt{\chi_\nu^2/\nu}}$$

follows a *t-distribution* with  $\nu$  degrees of freedom. Pdf of the t-distribution:

$$f_T(t) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\nu\pi}\Gamma(\frac{\nu}{2})} \left(1 + \frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}; -\infty < t < \infty$$

We will proceed as follows:

- A. Show  $T$  follows a  $t_\nu$  distribution.
- B. Let  $Y_1, \dots, Y_n$  be an i.i.d. sample from a  $N(\mu, \sigma^2)$  distribution. Let:

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

Show that this can be written as  $Z/\sqrt{W/(n-1)}$  where  $Z \sim N(0, 1)$  and  $W \sim \chi_{n-1}^2$ , and hence that  $T \sim t_{n-1}$ .

**Proof of A**

(Proof of A, continued)

B. Let  $Y_1, \dots, Y_n$  be an i.i.d. sample from a  $N(\mu, \sigma^2)$  distribution. Let:

$$T = \frac{\bar{Y} - \mu}{s/\sqrt{n}}$$

Show that this can be written as  $Z/\sqrt{W/(n-1)}$  where  $Z \sim N(0, 1)$  and  $W \sim \chi_{n-1}^2$ , and hence that  $T \sim t_{n-1}$ .

**Deriving 95% confidence intervals for  $\mu$**

Suppose  $Y_1, Y_2, \dots, Y_n$  is an i.i.d. sample drawn from a  $N(\mu, \sigma^2)$  population. Derive a 95% confidence interval for  $\mu$ .

```

#Write code to get sample, calculate 95% confidence interval
get.one.ci <- function(n){
  one.sample <- rnorm(n, mean = 2, sd = sqrt(4))
  ybar <- mean(one.sample)
  s <- sd(one.sample)
  lower <- ybar - qt(0.975,n-1)*s/sqrt(n)
  upper <- ybar + qt(0.975,n-1)*s/sqrt(n)
  ci <- c(lower,upper)
  return(ci)
}

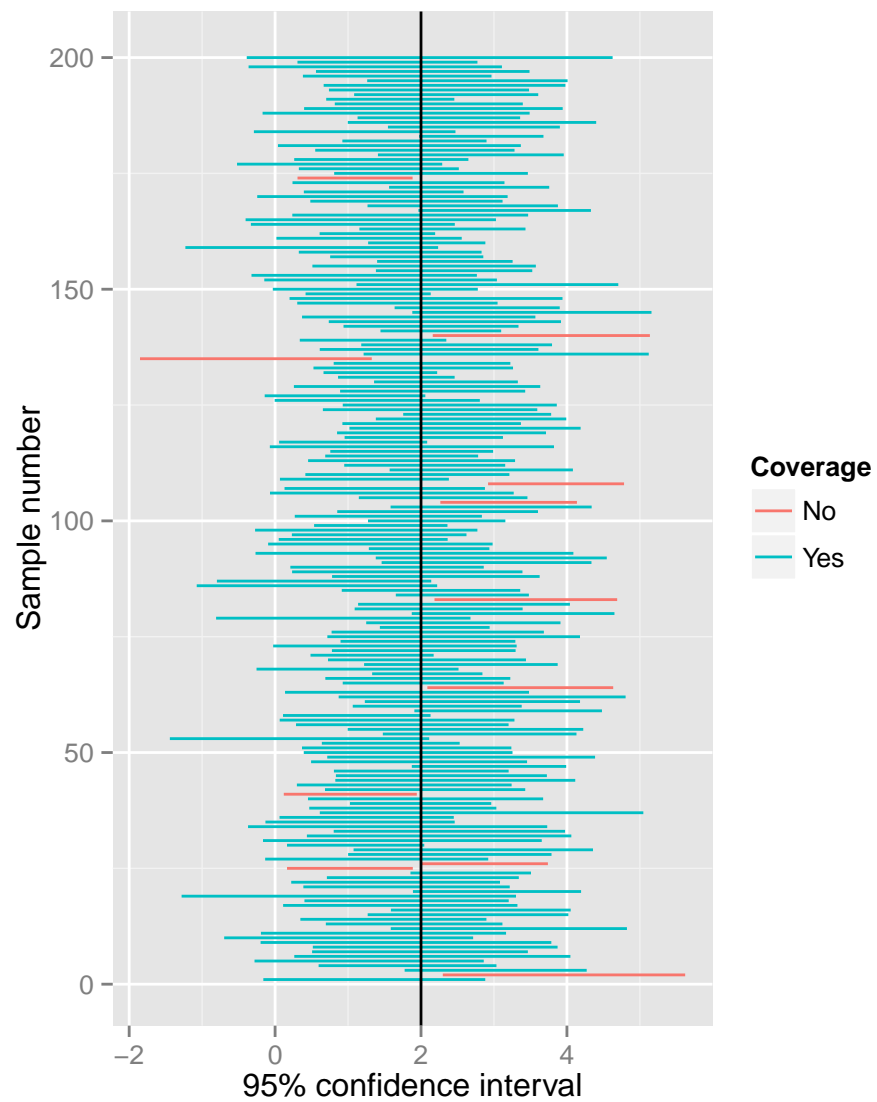
#Given a 95% confidence interval, and a value of the true mu, does the interval cover mu?
covers.mu <- function(ci,mu) {
  cover <- ifelse(ci[1] < mu & mu < ci[2], 'Yes', 'No')
  return(cover)
}

#Gather 200 samples and corresponding confidence intervals, and calculate coverage:
set.seed(24111)
many.ci <- replicate(200, get.one.ci(n=10), simplify='matrix')
df <- data.frame(t(many.ci))
df$Coverage <- apply(df, 1, covers.mu, mu=2)
df$Sample <- 1:nrow(df)
table(df$Coverage)/200

##
##      No      Yes
## 0.055 0.945

```

```
##Plot the results
library(ggplot2)
ggplot(data = df) +
  geom_segment(aes(x = X1, xend = X2, y = Sample, yend = Sample,color=Coverage)) +
  geom_vline(xintercept=2) + xlab('95% confidence interval') + ylab('Sample number')
```



#### Showing #4

If  $Y_1, Y_2, \dots, Y_n$  is an i.i.d. sample from a  $N(\mu_Y, \sigma_Y^2)$  distribution; and  $X_1, X_2, \dots, X_m$  is a sample drawn i.i.d from a  $N(\mu_X, \sigma_X^2)$  distribution; then the ratio of the sample variances scaled as follows,  $\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2}$ , follows an  $F_{m-1, n-1}$  distribution.

A. First, we need to prove that **in general**, if  $U \sim \chi_p^2$  and  $V \sim \chi_q^2$ , then  $W = \frac{U/p}{V/q} \sim F_{p,q}$  where:

$$f_W(w) = \frac{\Gamma\left(\frac{p+q}{2}\right)}{\Gamma\left(\frac{p}{2}\right)\Gamma\left(\frac{q}{2}\right)} \left(\frac{p}{q}\right)^{p/2} w^{p/2-1} \left(1 + \frac{p}{q}w\right)^{-\left(\frac{p+q}{2}\right)}; w > 0$$

B. Then, we need to prove that  $\frac{S_X^2/\sigma_X^2}{S_Y^2/\sigma_Y^2} \equiv \frac{\chi_{m-1}^2/(m-1)}{\chi_{n-1}^2/(n-1)}$

Proof of A:



Proof of B:

**Usage: hypothesis testing for equality of two population variances**

Suppose we have  $X_1, X_2, \dots, X_m$  drawn i.i.d.  $\sim N(\mu_X, \sigma_X^2)$  and  $Y_1, Y_2, \dots, Y_n \sim N(\mu_Y, \sigma_Y^2)$ . We are interested in testing whether the two population variances are equal, e.g.:

$$H_0 : \sigma_X^2 = \sigma_Y^2$$

$$H_a : \sigma_X^2 \neq \sigma_Y^2$$

How can we derive a test for these hypotheses?

## The central limit theorem

One of the most important theorems in statistics, the central limit theorem (CLT) guarantees normality of  $\bar{Y}$  for large  $n$ , no matter what distribution the individual  $Y_i$  themselves came from.

Here is the theorem in all its glory:

Let  $Y_1, Y_2, \dots, Y_n$  be i.i.d. random variables with  $E(Y_i) = \mu$  and  $Var(Y_i) = \sigma^2 < \infty$ . *Note that no assumptions are made about normality of the individual  $Y_i$ !* Let:

$$U_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \sqrt{n} \left( \frac{\bar{Y} - \mu}{\sigma} \right).$$

Then, as  $n \rightarrow \infty$ ,  $U_n \rightarrow_d N(0, 1)$ .

The statement  $\rightarrow_d$  means “converges in distribution.” Essentially what this means is that, as  $n \rightarrow \infty$ ,

$$P(U_n \leq u) \rightarrow \int_{-\infty}^u \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt,$$

i.e. the CDF of a standard normal.

---

A couple points of clarification:

No matter what distribution the  $Y_i$  come from:

1. No matter what size  $n$ ,  $E(\bar{Y}) = \mu$  and  $Var(\bar{Y}) = \sigma^2/\sqrt{n}$ . This is Chapter 4 stuff.
2. What the CLT gives us is *normality* of the  $\bar{Y}$  for large  $n$ .

Before proving this, let's investigate the CLT via simulations. We'll take repeated samples of  $EXP(\beta = 5)$  random variables, of various sizes. Note from here:

- $\mu = E(Y_i) = \beta = 5$
- $\sigma^2 = Var(Y_i) = \beta^2 = 25$
- $E(\bar{Y}) = \mu = 5$  for all  $n$
- $Var(\bar{Y}) = \sigma^2/n = 25/n$  for all  $n$
- $\bar{Y}$  are normal for *large  $n$  only*

```

get.one.ybar <- function(n){
  one.sample <- rexp(n, rate = 1/5)
  ybar <- mean(one.sample)
  return(ybar)
}
set.seed(12345)
many.ybar.n2 <- replicate(1000,get.one.ybar(n=2))
many.ybar.n5 <- replicate(1000,get.one.ybar(n=5))
many.ybar.n20 <- replicate(1000,get.one.ybar(n=20))
many.ybar.n50 <- replicate(1000,get.one.ybar(n=50))
df <- data.frame(many.ybar.n2,many.ybar.n5,many.ybar.n20,many.ybar.n50)
apply(df,2,mean) #Should all be ~5:

```

```

## many.ybar.n2 many.ybar.n5 many.ybar.n20 many.ybar.n50
##      4.928742      5.009634      5.001159      4.974126

```

```

apply(df,2,var) #Should be decreasing:

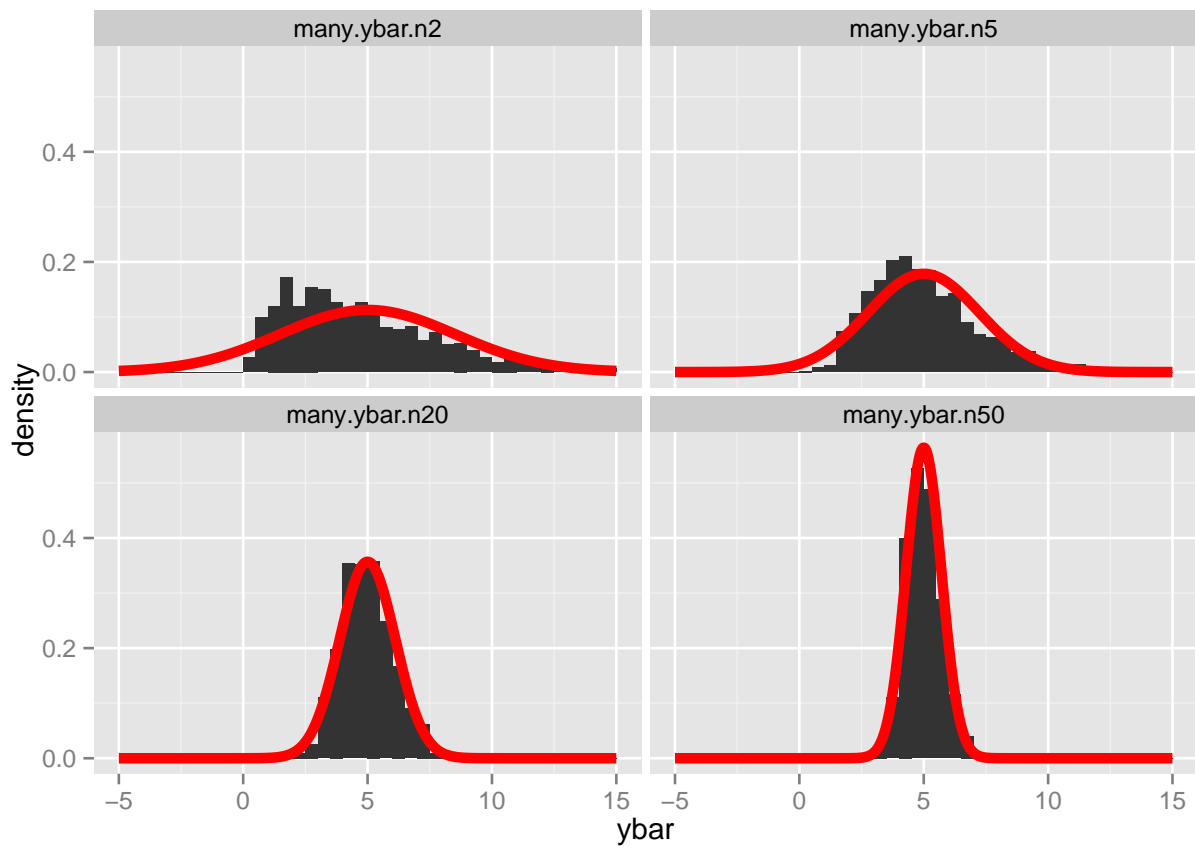
```

```

## many.ybar.n2 many.ybar.n5 many.ybar.n20 many.ybar.n50
##      11.8196198      4.6614922      1.1739804      0.4983923

```

```
library(tidyr)
df2 <- gather(df, key = 'SampleSize', value = 'ybar')
xseq <- seq(-5,15,l=1000)
df2$xseq <- rep(xseq,4)
df2$yseq <- c(dnorm(xseq, mean = 5, sd = sqrt(25/2)),
              dnorm(xseq, mean = 5, sd = sqrt(25/5)),
              dnorm(xseq, mean = 5, sd = sqrt(25/20)),
              dnorm(xseq, mean = 5, sd = sqrt(25/50)))
ggplot(data = df2) +
  geom_histogram(aes(x = ybar, y = ..density..),binwidth =.5) +
  geom_line(aes(x = xseq, y = yseq),color='red',size=2) +
  facet_wrap(~SampleSize) + xlim(c(-5,15))
```



### Proof of the CLT: preliminaries

To prove the CLT, we will use the method of MGFs. Before we embark, recall a couple important definitions and facts from calculus:

*Definition:* A function  $f(n)$  is  $o(n)$  (“little oh of  $n$ ”) if it goes to 0 faster than  $n$  does. Specifically, if  $\lim_{n \rightarrow \infty} n f(n) \rightarrow 0$ .

Examples:  $f(n) = \frac{1}{n^2} = o(n)$ ;  $f(n) = \frac{1}{\sqrt{n}} \neq o(n)$ .

We also need the following facts:

- *Fact #1, from calculus:* For any  $t$ ,  $(1 + \frac{t}{n} + o(n)) \rightarrow e^t$ .
- *Fact #2, from earlier this semester:* Let  $M_Y(t)$  be the MGF of  $Y$ ; then  $M_{aY+b}(t) = e^{bt} M_Y(at)$ .
- *Fact #3, from earlier this semester:* If  $Y_1, Y_2, \dots, Y_n$  are i.i.d. and  $S_n = \sum_{i=1}^n Y_i$ , then  $M_{S_n}(t) = M_Y(t)^n$ .

Given these facts, here is what we want to prove:

**The CLT, technically stated:** Let  $Y_1, Y_2, \dots, Y_n$  be an i.i.d. sample with  $|E(Y)| = |\mu| < \infty$  and  $0 < E(Y^2) < \infty$ . Let:

$$U_n = \frac{\bar{Y} - \mu}{\sigma/\sqrt{n}} = \frac{\frac{1}{n} \sum_{i=1}^n (Y_i - \mu)}{\sigma/\sqrt{n}} = \frac{\sum_{i=1}^n (Y_i - \mu)}{\sqrt{n}\sigma} = \frac{\sum_{i=1}^n X_i}{\sqrt{n}\sigma},$$

where  $X_i = (Y_i - \mu)$ . Show that  $M_{U_n}(t) \rightarrow e^{t^2/2}$  as  $n \rightarrow \infty$ , where  $e^{t^2/2}$  is the MGF of a  $N(0, 1)$  distribution; hence showing that  $U_n \rightarrow_d N(0, 1)$ .

**PROOF:**

## Proof of CLT, continued