Unit 2: Foundations for Inference

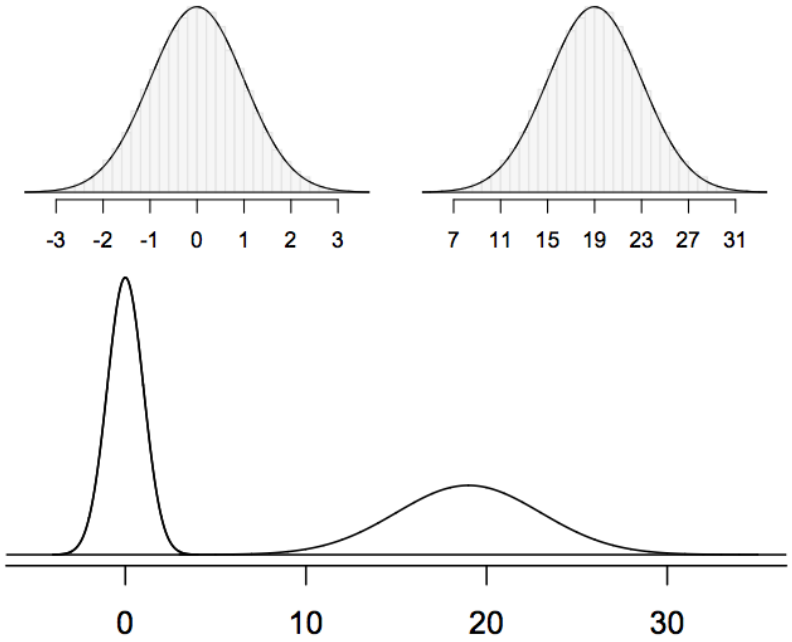
3. The Normal Distribution

10/18/2017

Different Normal Distributions

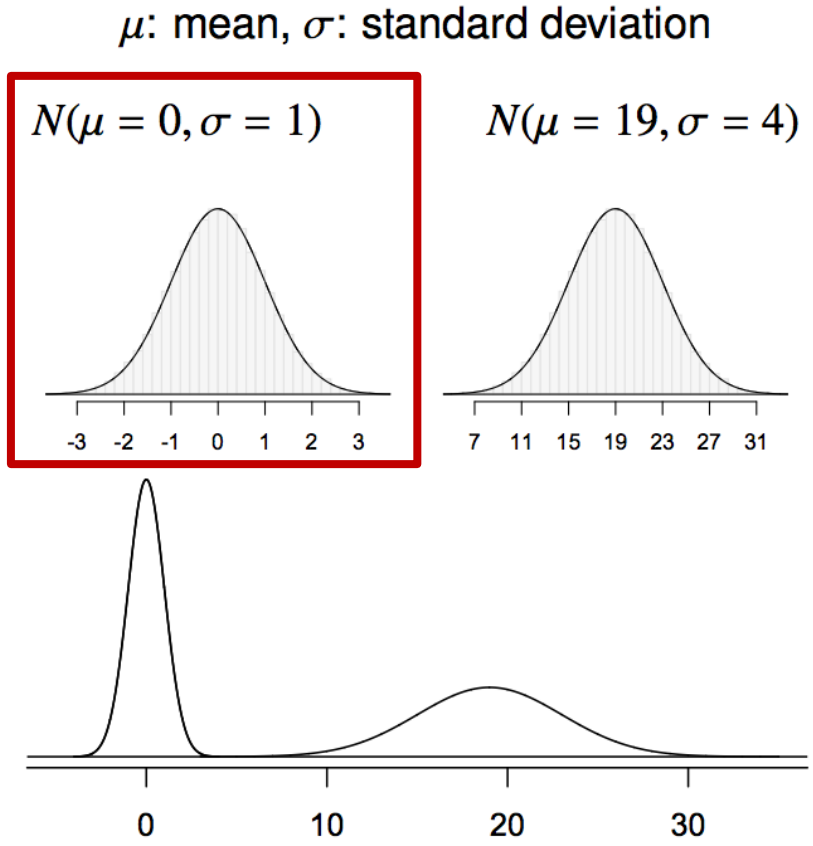
μ: mean, σ: standard deviation

$N(\mu = 0, \sigma = 1)$ $N(\mu = 19, \sigma = 4)$



Different Normal Distributions

Standard Normal Distribution



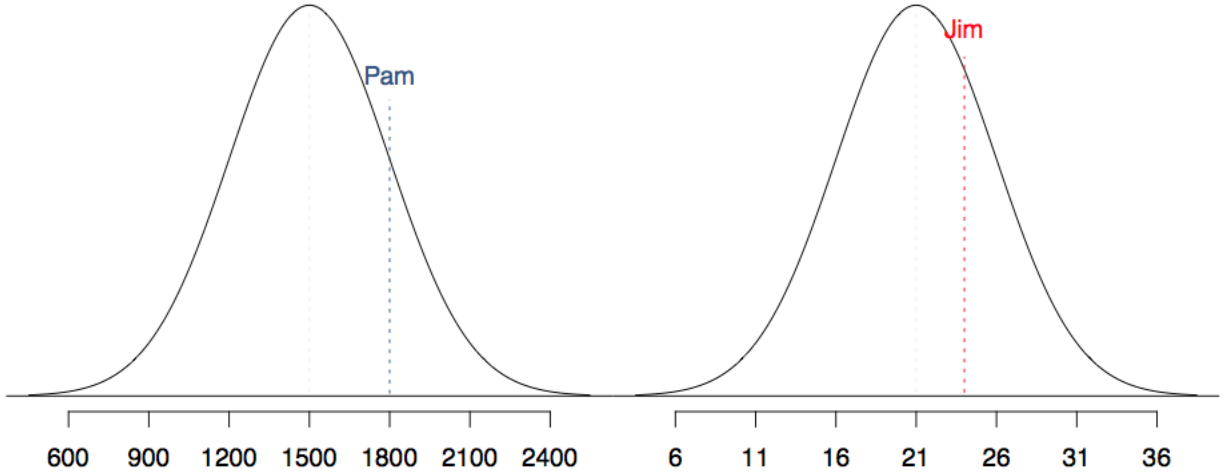
Comparing samples from two normal distributions

A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers:
Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

Comparing samples from two normal distributions

A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers:
Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.
ACT scores are distributed nearly normally with mean 21 and standard deviation 5.

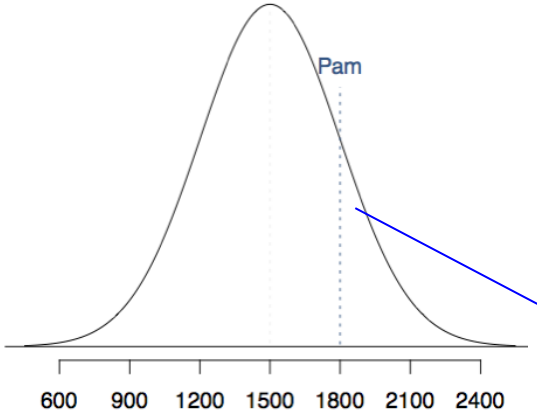


Comparing samples from two normal distributions

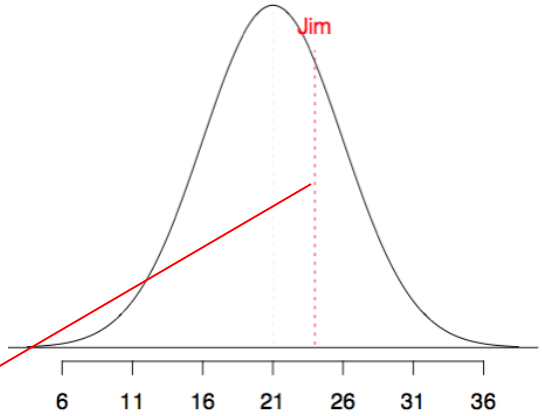
A college admissions officer wants to determine which of the two applicants scored better on their standardized test with respect to the other test takers:
Pam, who earned an 1800 on her SAT, or Jim, who scored a 24 on his ACT?

SAT scores are distributed nearly normally with mean 1500 and standard deviation 300.
ACT scores are distributed nearly normally with mean 21 and standard deviation 5.

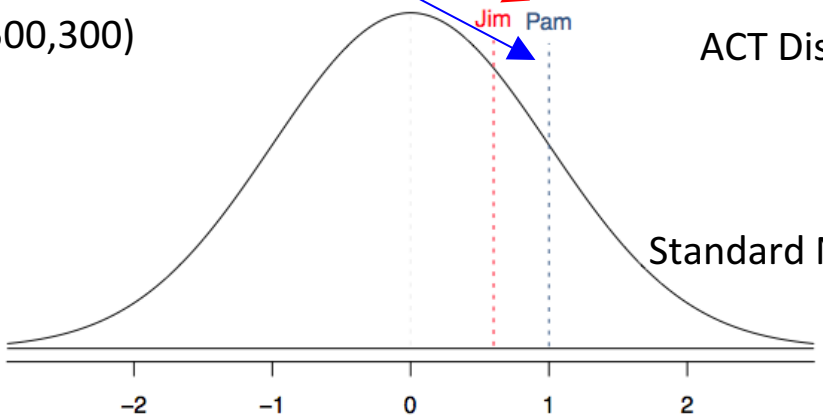
We can map different Normal Distributions onto the Standard Normal



SAT Distribution: N(1500,300)



ACT Distribution: N(21,5)



Standard Normal: N(0,1)

Z-score: Number of Standard Deviations above the mean

These are called standardized scores, or Z-scores.

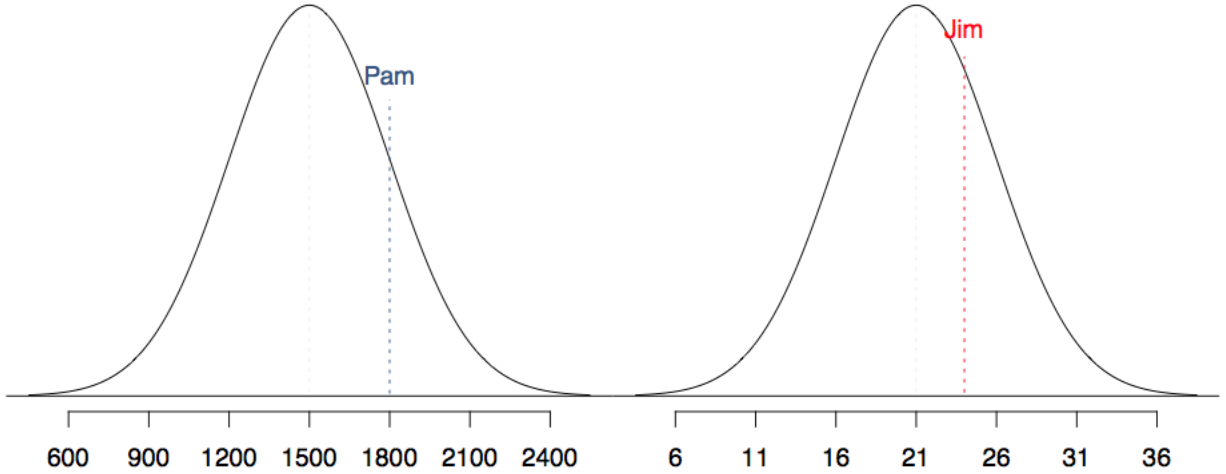
- Z-score of an observation is the number of standard deviations it falls above or below the mean.

  Z = (observation - mean) / SD

Comparing samples from two normal distributions

We can't just compare these two raw scores. But, we *can* compare how many standard deviations beyond the mean each observation is.

- Pam's score is (1800 - 1500) / 300 = 1 standard deviation above the mean.
- Jim's score is (24 - 21) / 5 = 0.6 standard deviations above the mean.



Z-score: Number of Standard Deviations above the mean

These are called standardized scores, or Z-scores.

- Z-score of an observation is the number of standard deviations it falls above or below the mean.

  Z = (observation - mean) / SD

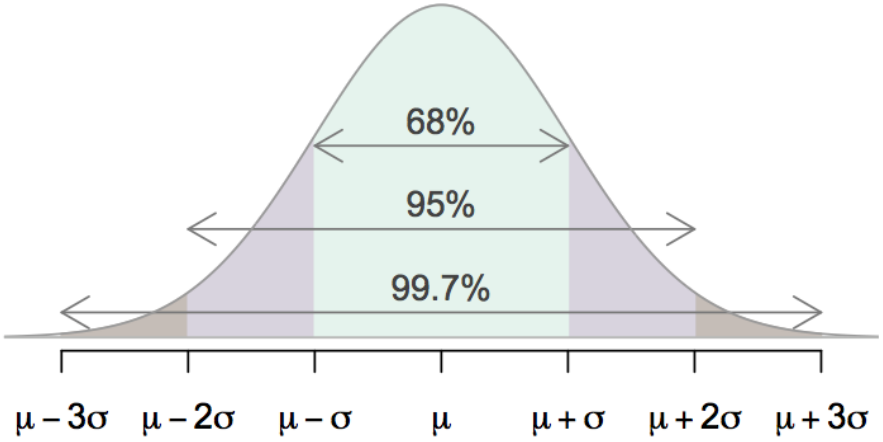
Observations that are more than 2 SD away from the mean ($|Z| > 2$) are usually considered unusual.

Z scores are defined for distributions of any shape, but only when the distribution is normal can we use Z scores to calculate percentiles.

The 68-95-99.7 Rule

For nearly normally distributed data,
about 68% falls within 1 SD of the mean,
about 95% falls within 2 SD of the mean,
about 99.7% falls within 3 SD of the mean.

It is possible for observations to fall 4, 5, or more standard deviations away from the mean, but these occurrences are very rare if the data are nearly normal.



Practice Question 1: Quality Control

At the Heinz ketchup factory, the amount of ketchup that goes into the bottle is supposed to be normally distributed with mean 36 oz. and standard deviation 0.11 oz.

Once every 30 minutes a bottle is selected from the production line, and its contents are noted precisely. If the amount of ketchup in the bottle is below 35.8 oz. or above 36.2 oz., then the bottle fails the quality control inspection.

What percent of bottles have less than 35.8 ounces of ketchup?

$Z(35.8) = (35.8 - 36)/.11 \sim -1.82$

Since ~95% of the distribution falls within 2SD on either side of the mean, we should expect it to be a little bit more than 2.5%

Practice Question 2: Quality Control

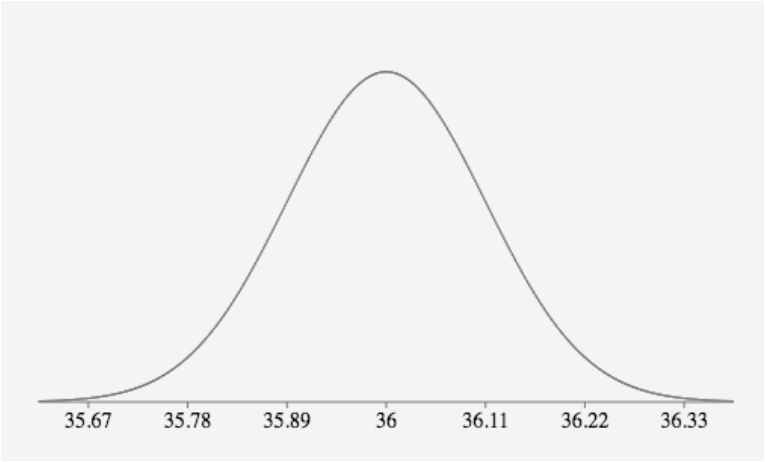
What percent of bottles pass the quality control inspection?

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%

Practice Question 2: Quality Control

What percent of bottles pass the quality control inspection?

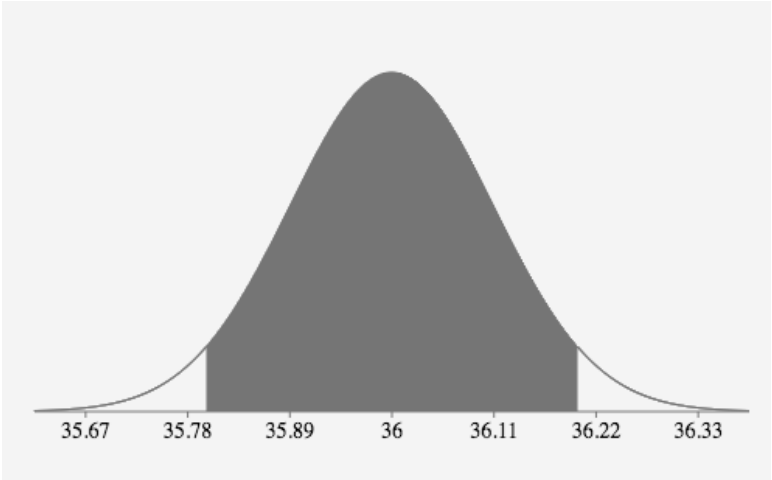
(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%



Practice Question 2: Quality Control

What percent of bottles pass the quality control inspection?

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%



Practice Question 2: Quality Control

What percent of bottles pass the quality control inspection?

(a) 1.82%
(b) 3.44%
(c) 6.88%
(d) 93.12%
(e) 96.56%

Aside: Don't worry about probability tables. We live in 2017

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

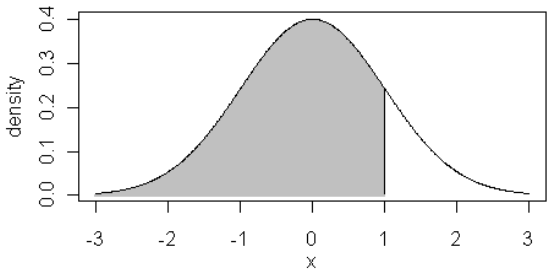
use *pnorm(Z)*

Aside: Don't worry about probability tables. We live in 2017

Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5

use *pnorm(Z)*

pnorm(observation, mean, SD)



Practice Question 3: Which of the following is **false**

1. Majority of Z-scores in a right skewed distribution are negative.
2. In skewed distributions the Z-score of the mean might be different than 0.
3. For a normal distribution, IQR is less than 2 x SD.
4. Z-scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Practice Question 3: Which of the following is **false**

1. Majority of Z-scores in a right skewed distribution are negative.
2. **In skewed distributions the Z-score of the mean might be different than 0.**
3. For a normal distribution, IQR is less than 2 x SD.
4. Z-scores are helpful for determining how unusual a data point is compared to the rest of the data in the distribution.

Key ideas

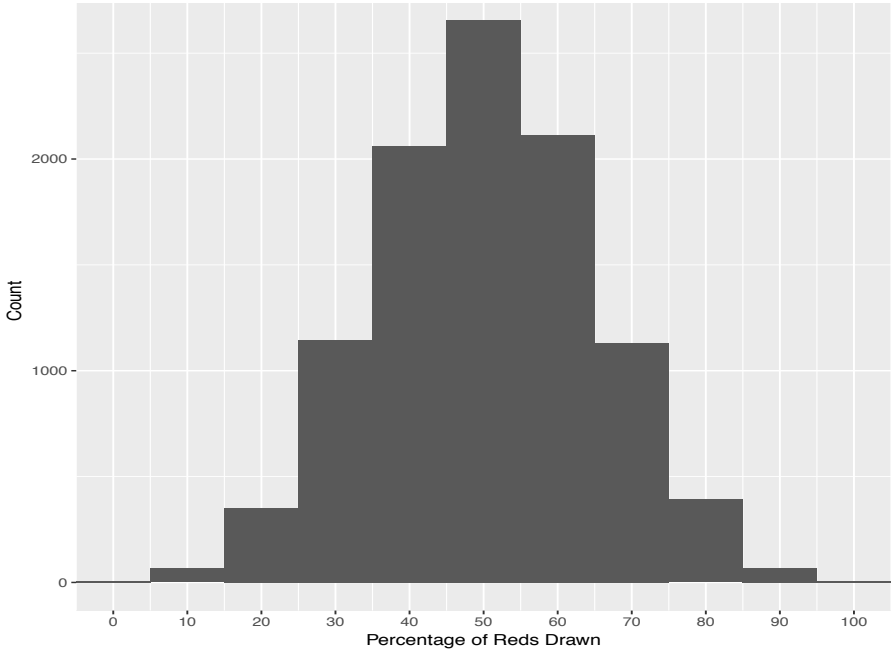
1. We are really thinking about three distributions: the sample, the population, and the test statistic

2. We can use Z-scores to compare points on two different normal distributions

Inference For Proportions

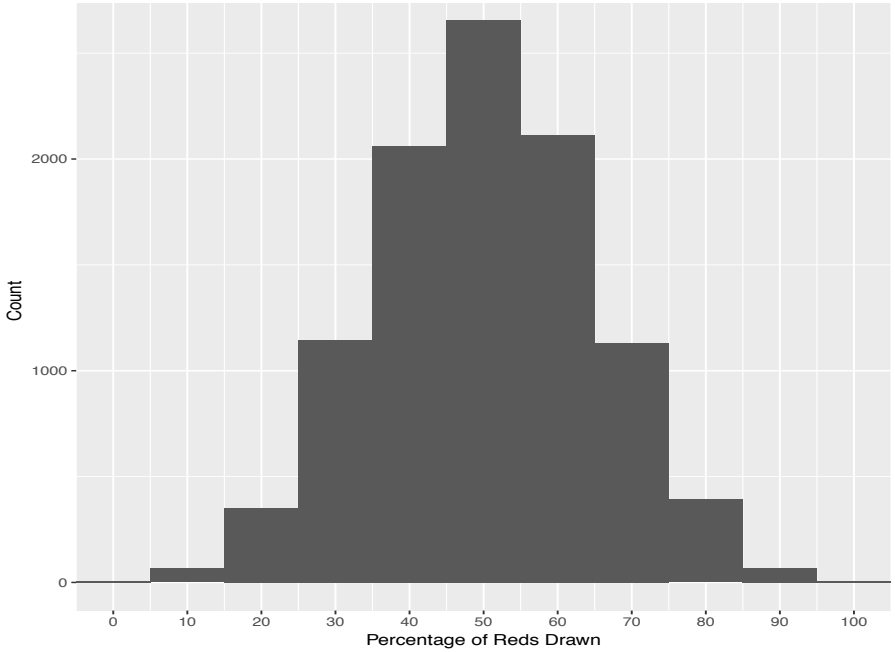
- Sometimes we're interested in knowing the *proportion* of a variable in the population. (e.g. What percent of marbles in the box are red?)
- Step 1. Take a sample, calculate the proportion in the sample
  - I pulled 10 marbles out of the box, and 90% were red.
- Step 2. Posit a null hypothesis about the population
  - Null Hypothesis (H_0): "50% of the marbles in the box are red"
- Step 3. Then, calculate a sampling distribution *assuming* the null is true
  - If a box *does have* 50% red marbles, and I pull 10 out, how often do expect to see 10% red in my sample? 20% red? 30% red? etc.

Note: The exact shape of this sampling distribution depends on (1) population proportion, and (2) the sample size



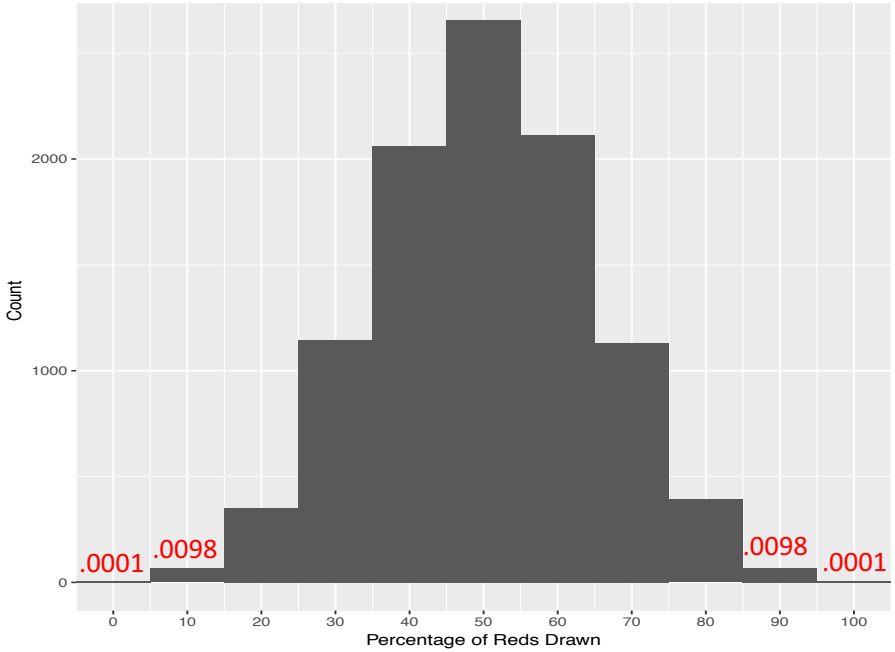
Inference For Proportions

- Sometimes we're interested in knowing the *proportion* of a variable in the population. (e.g. What percent of marbles in the box are red?)
- Step 1. Take a sample, calculate the proportion in the sample
  - I pulled 10 marbles out of the box, and 90% were red.
- Step 2. Posit a null hypothesis about the population
  - Null Hypothesis (H_0): "50% of the marbles in the box are red"
- Step 3. Then, calculate a sampling distribution *assuming* the null is true
  - If a box *does have* 50% red marbles, and I pull 10 out, how often do expect to see 10% red in my sample? 20% red? 30% red? etc.
- Step 4. See where our sample falls in the sampling distribution. If the null were true, how often would we expect to see results this extreme or more?



Inference For Proportions

- Sometimes we're interested in knowing the *proportion* of a variable in the population. (e.g. What percent of marbles in the box are red?)
- Step 1. Take a sample, calculate the proportion in the sample
  - I pulled 10 marbles out of the box, and 90% were red.
- Step 2. Posit a null hypothesis about the population
  - Null Hypothesis (H_0): "50% of the marbles in the box are red"
- Step 3. Then, calculate a sampling distribution *assuming* the null is true
  - If a box *does have* 50% red marbles, and I pull 10 out, how often do expect to see 10% red in my sample? 20% red? 30% red? etc.
- Step 4. See where our sample falls in the sampling distribution. If the null were true, how often would we expect to see results this extreme or more?
  - If the box did have 50% marbles, we'd get 90% red in a sample of 10 (or more extreme) only 1.98% of the time



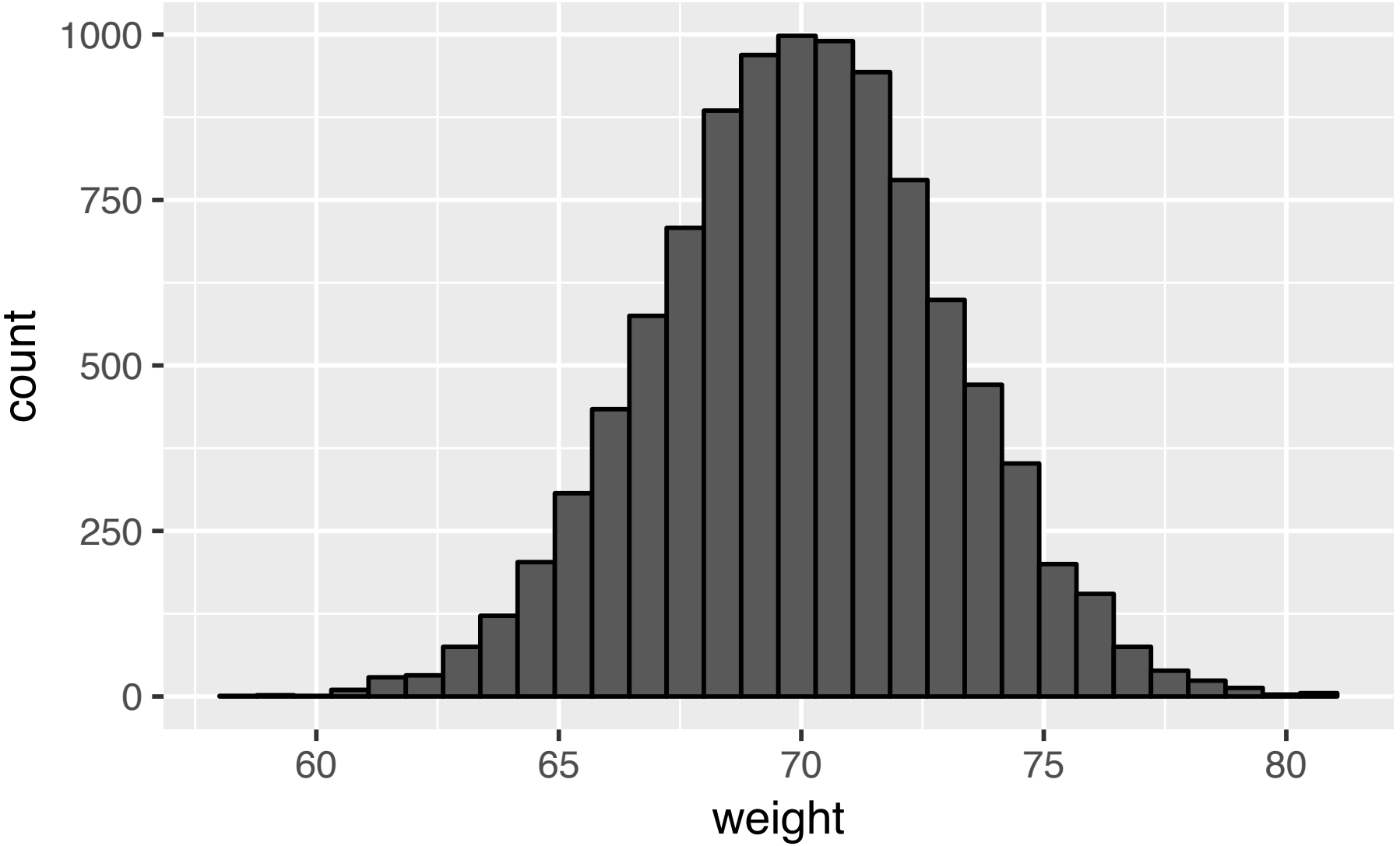
Inference For Means

- Sometimes we're interested in knowing the *mean* of a variable in the population.
  - What's the mean weight of the apples in my orchard?
- Step 1. Take a sample, calculate the mean in the sample
  - I weighed 10 apples and the mean weight was 68 grams.
- Step 2. Posit a null hypothesis about the population
  - The mean weight of apples in my orchard is 70 grams with a SD of 3 grams (same as all apples of this type)
- Step 3. Calculate a "sampling distribution" *assuming* the null is true
  - If the apples in the orchard had a mean weight of 70 grams and SD of 3 grams, how often would I get a sample of 50 apples with a mean of 60? 61? 62? 63? etc.
- Step 4. See where our sample falls in the sampling distribution. If the null were true, how often would we expect to see results this extreme or more?

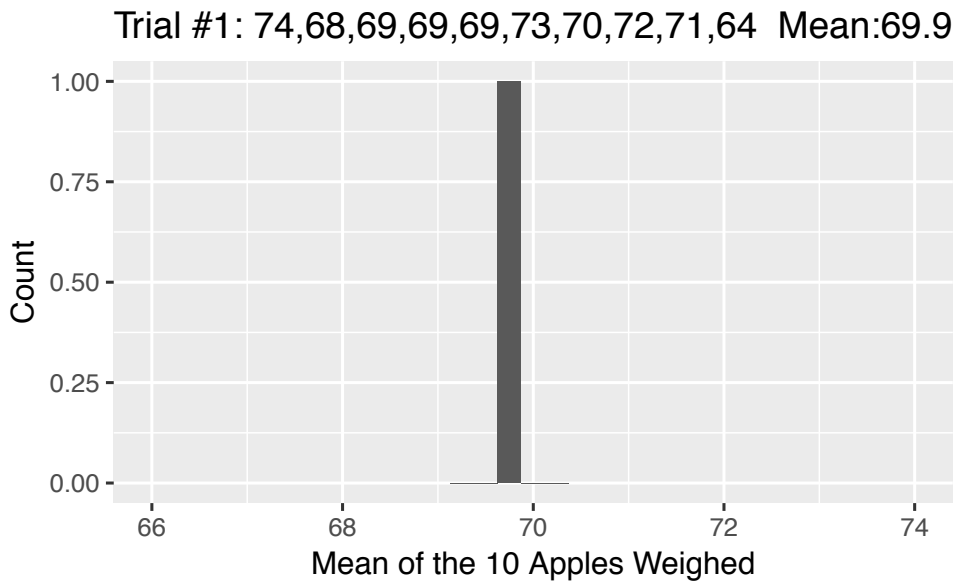
Inference For Means

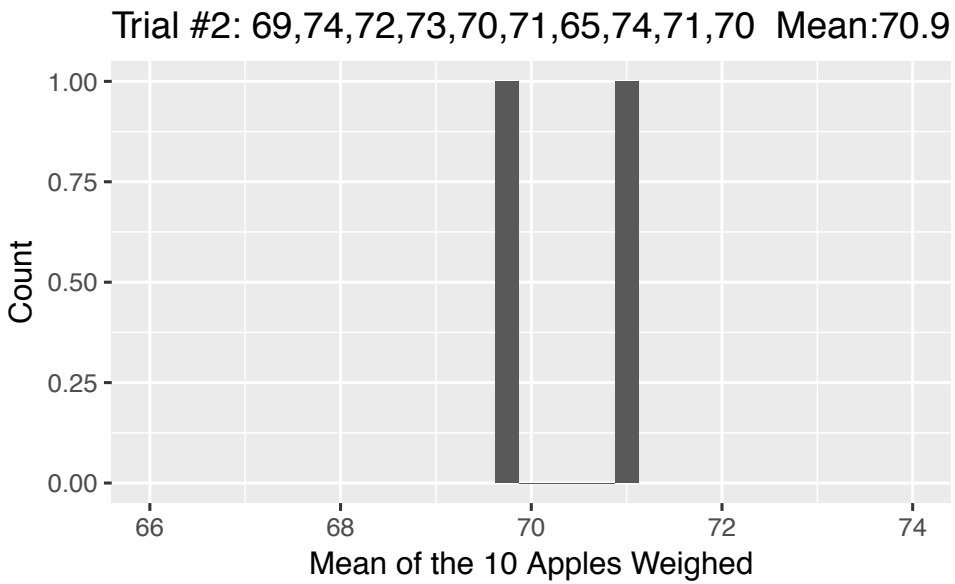
- Sometimes we're interested in knowing the *mean* of a variable in the population.
  - What's the mean weight of the apples in my orchard?
- Step 1. Take a sample, calculate the mean in the sample
  - I weighed 10 apples and the mean weight was 68 grams.
- Step 2. Posit a null hypothesis about the population
  - The mean weight of apples in my orchard is 70 grams with a SD of 3 grams (same as all apples of this type)
- Step 3. Calculate a "sampling distribution" *assuming* the null is true
  - If the apples in the orchard had a mean weight of 70 grams and SD of 3 grams, how often would I get a sample of 50 apples with a mean of 60? 61? 62? 63? etc.
- Step 4. See where our sample falls in the sampling distribution. If the null were true, how often would we expect to see results this extreme or more?
  - If the apples in the orchard really did have mean 70 and SD 3, I'd expect to pick 10 apples with a mean of 68 (or more extreme) only 3.5% of the time

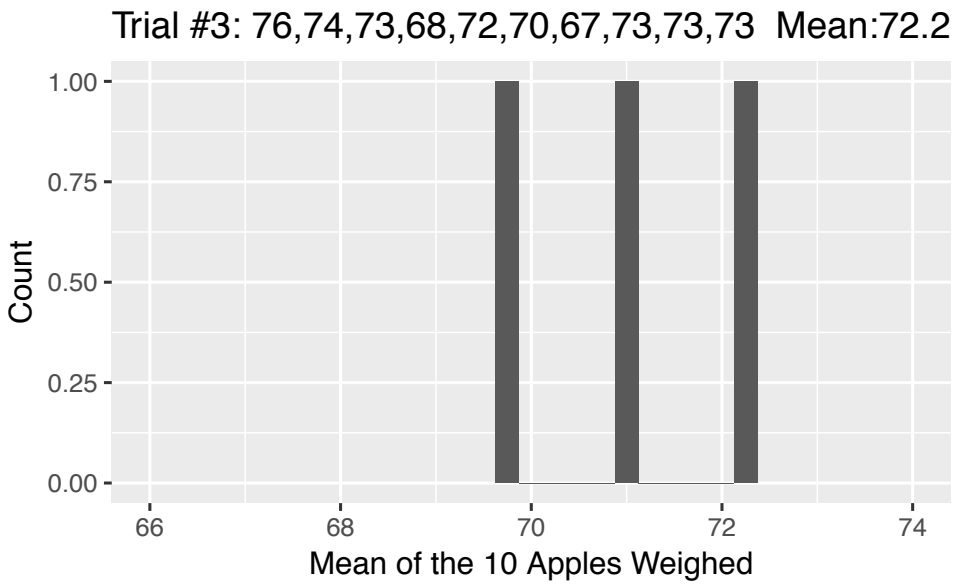
Hypothesized orchard population:

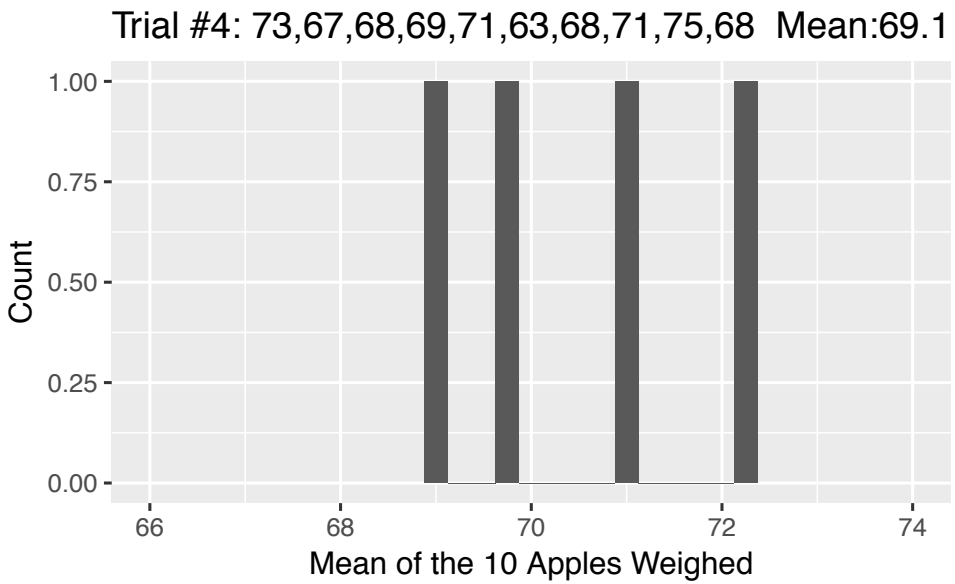


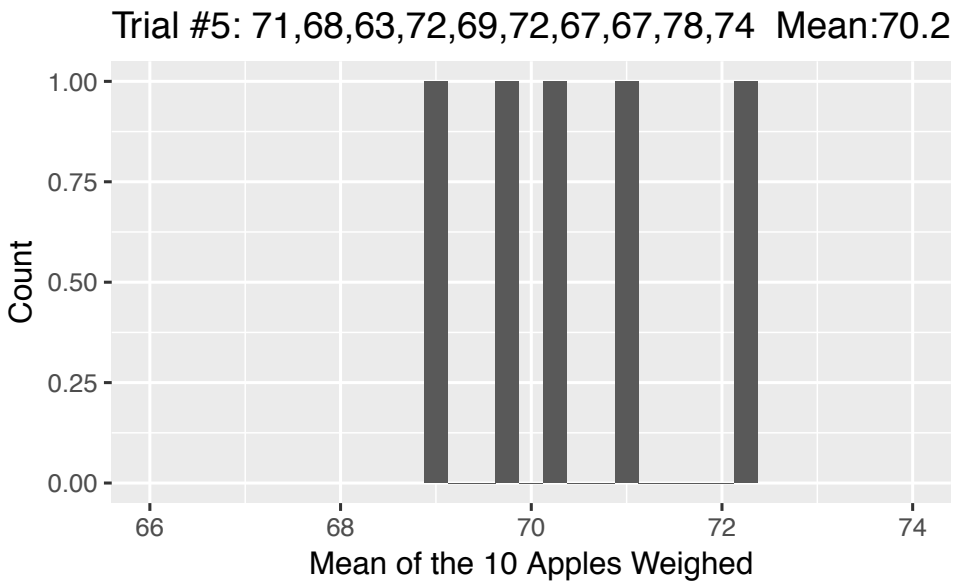
Normally Distributed, Mean: 70, Standard Deviation: 3

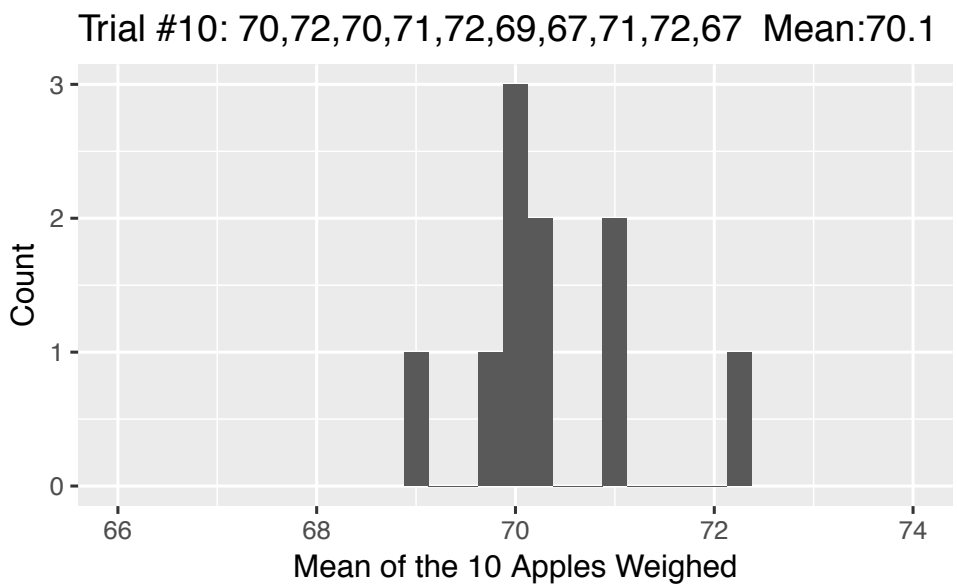

Trial #1: 74,68,69,69,69,73,70,72,71,64 Mean:69.9
1.00
0.75
0.50
0.25
0.00
Count
66
68
70
72
74
Mean of the 10 Apples Weighed

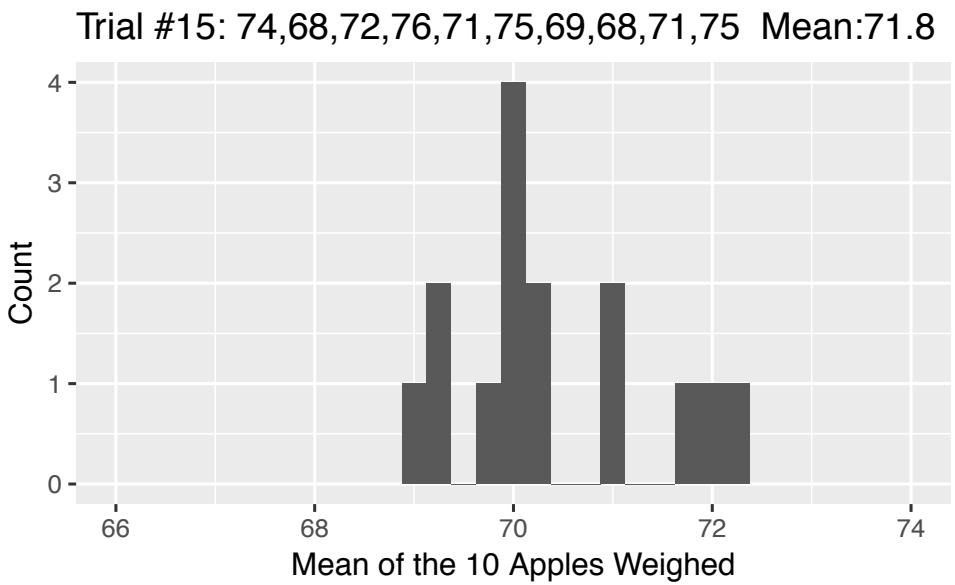

Trial #2: 69,74,72,73,70,71,65,74,71,70 Mean:70.9
1.00
0.75
0.50
0.25
0.00
Count
66
68
70
72
74
Mean of the 10 Apples Weighed

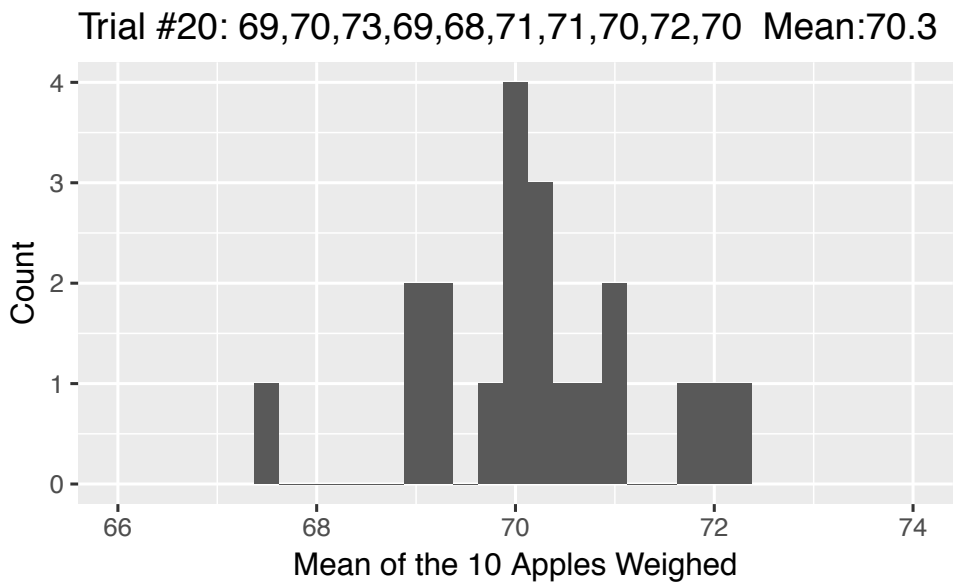

Trial #3: 76,74,73,68,72,70,67,73,73,73 Mean:72.2
1.00
0.75
0.50
0.25
0.00
Count
66
68
70
72
74
Mean of the 10 Apples Weighed

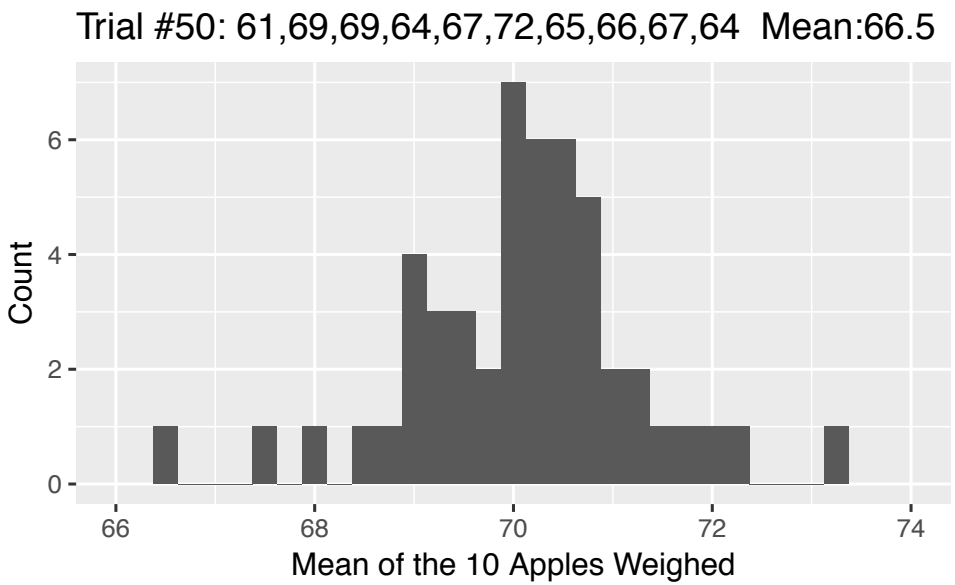

Trial #4: 73,67,68,69,71,63,68,71,75,68 Mean:69.1
1.00
0.75
0.50
0.25
0.00
Count
66
68
70
72
74
Mean of the 10 Apples Weighed

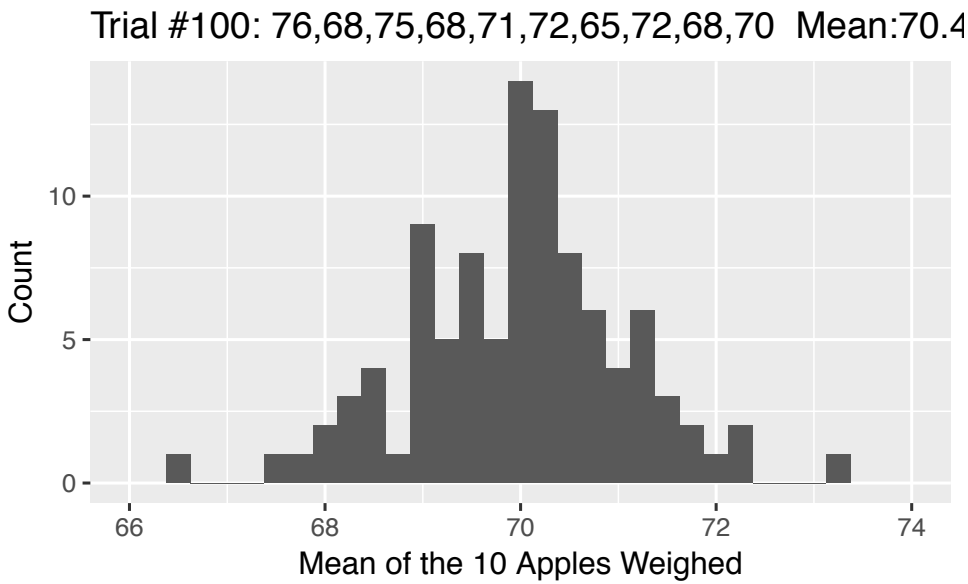

Trial #5: 71,68,63,72,69,72,67,67,78,74 Mean:70.2
1.00
0.75
0.50
0.25
0.00
Count
66
68
70
72
74
Mean of the 10 Apples Weighed

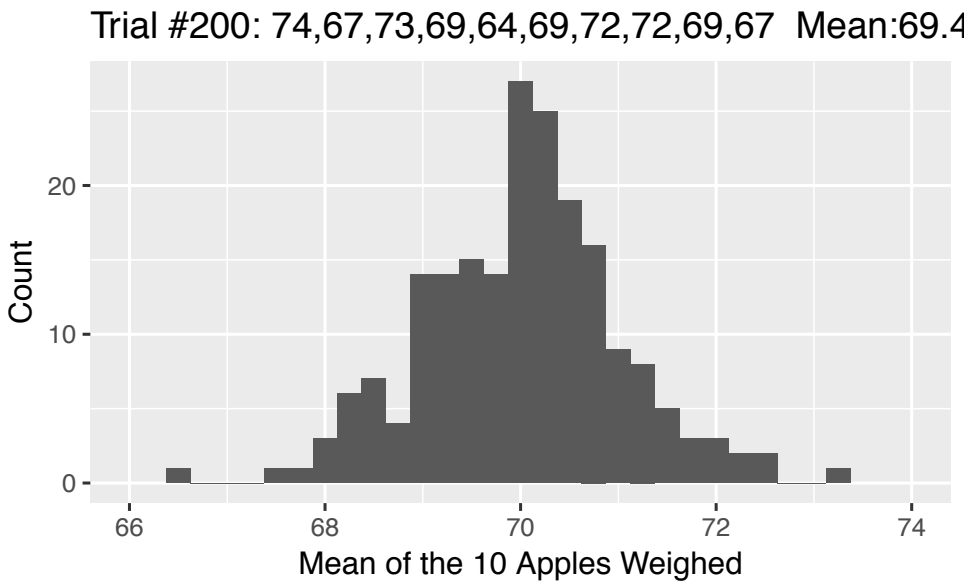

Trial #10: 70,72,70,71,72,69,67,71,72,67 Mean:70.1
Count
3
2
1
0
66
68
70
72
74
Mean of the 10 Apples Weighed

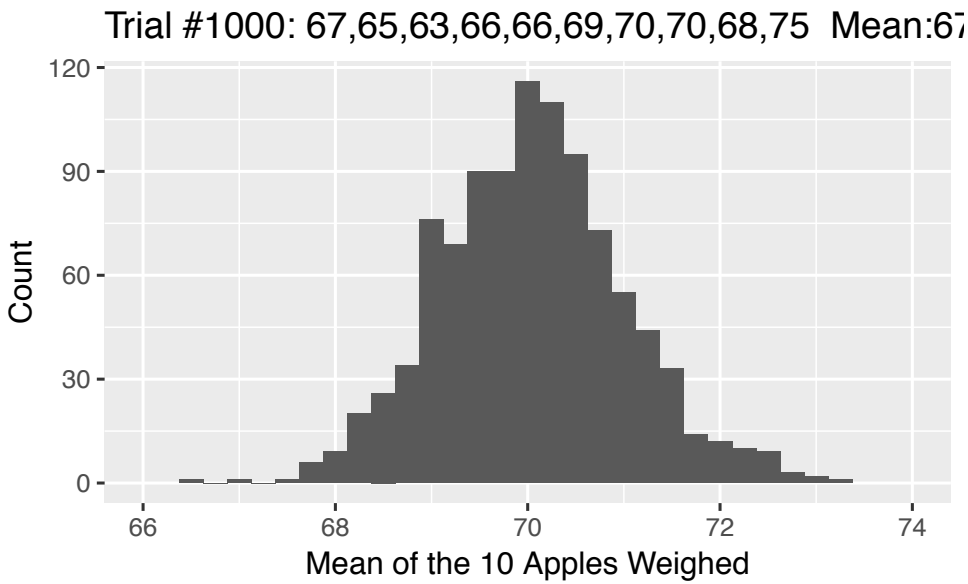

Trial #15: 74,68,72,76,71,75,69,68,71,75 Mean:71.8
Count
0
1
2
3
4
66
68
70
72
74
Mean of the 10 Apples Weighed

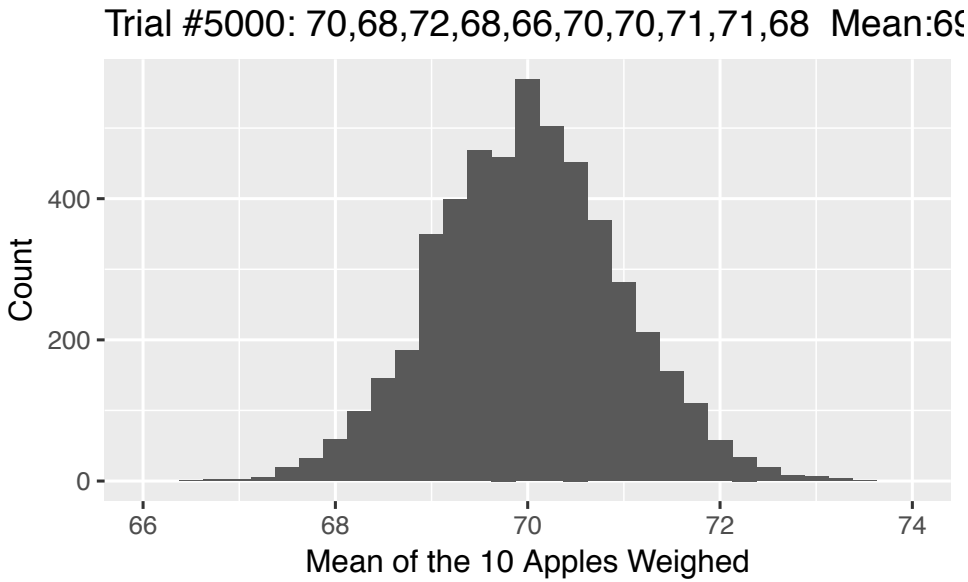

Trial #20: 69,70,73,69,68,71,71,70,72,70 Mean:70.3
Count
0
1
2
3
4
66
68
70
72
74
Mean of the 10 Apples Weighed


Trial #50: 61,69,69,64,67,72,65,66,67,64 Mean:66.5
Count
0
2
4
6
66
68
70
72
74
Mean of the 10 Apples Weighed

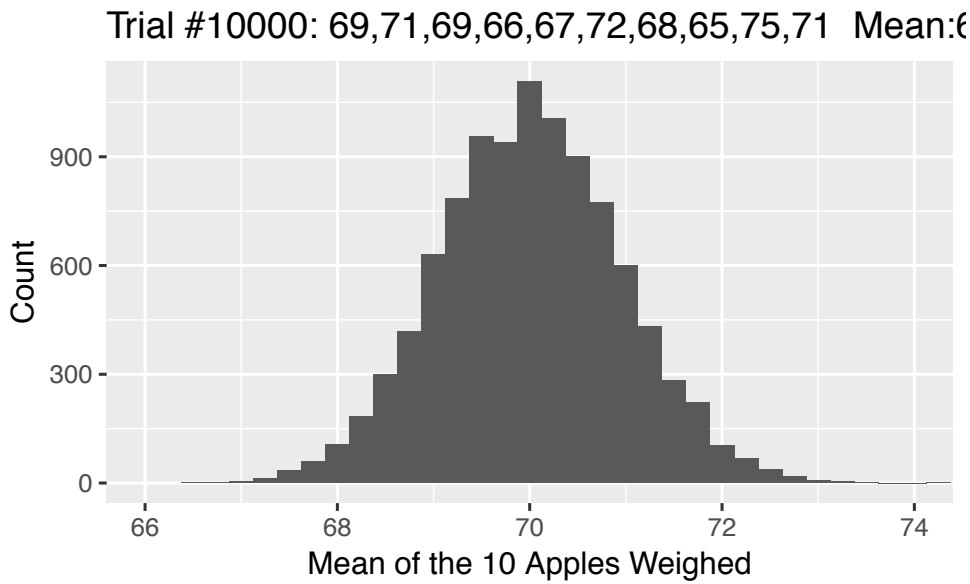

Trial #100: 76,68,75,68,71,72,65,72,68,70 Mean:70.4
Count
10
5
0
66
68
70
72
74
Mean of the 10 Apples Weighed


Trial #200: 74,67,73,69,64,69,72,72,69,67 Mean:69.4
Count
20
10
0
66
68
70
72
74
Mean of the 10 Apples Weighed


Trial #1000: 67,65,63,66,66,69,70,70,68,75 Mean:67
120
90
60
30
0
Count
66
68
70
72
74
Mean of the 10 Apples Weighed


Trial #5000: 70,68,72,68,66,70,70,71,71,68 Mean:69
Count
400
200
0
66
68
70
72
74
Mean of the 10 Apples Weighed

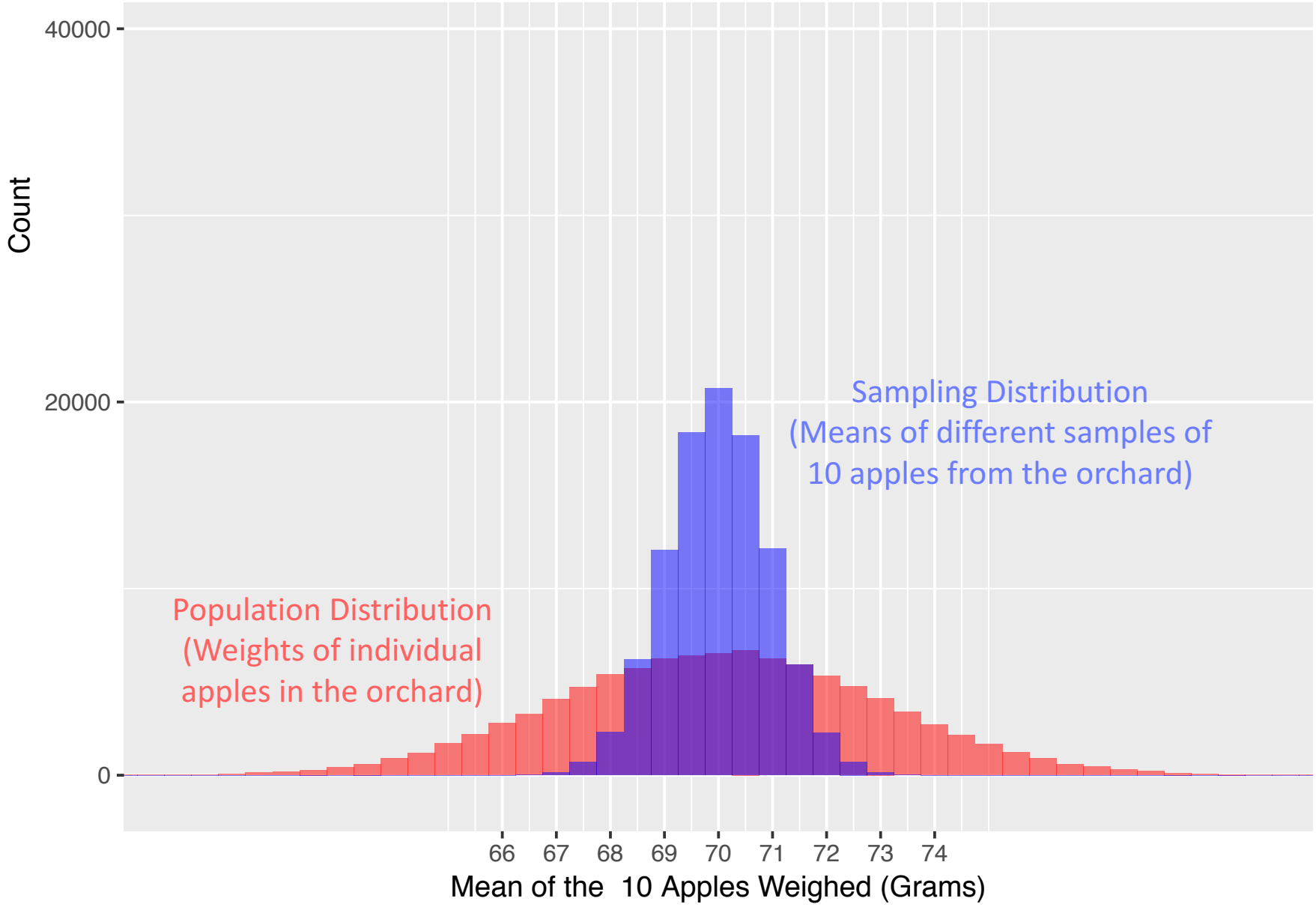
This is our Sampling distribution.



It shows how likely it is to see each different possible mean in a sample of 10 apples. (Assuming the samples were drawn from a population with mean 70 and SD 30)

In real life we will only draw ONE sample of 10 from the orchard. We then compare our sample to this sampling distribution to see how unlikely it is.

Note: Our sampling distribution is narrower than the underlying population distribution

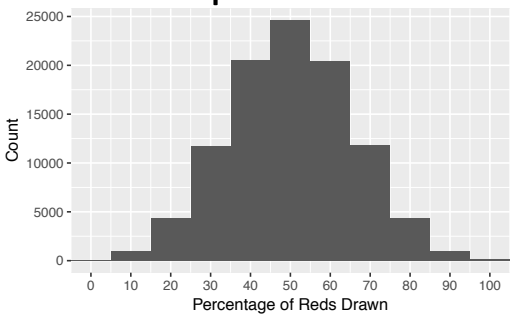


Remember some things we learned about proportions…

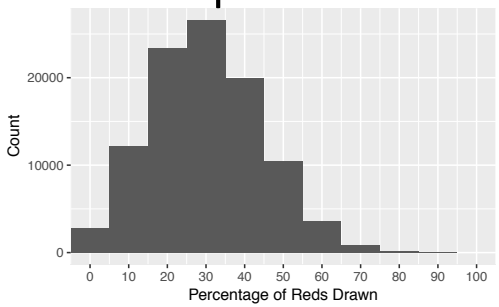
Remember some things we learned about proportions…

1. The farther the population proportion is from 0.5, the more skewed the sampling distribution is.

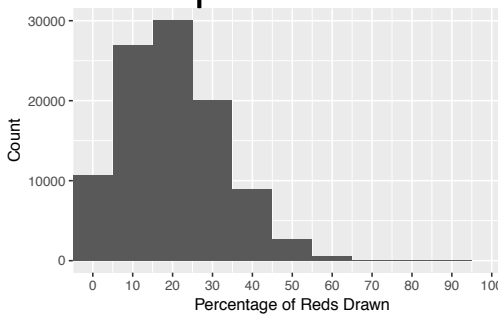
p = 0.5



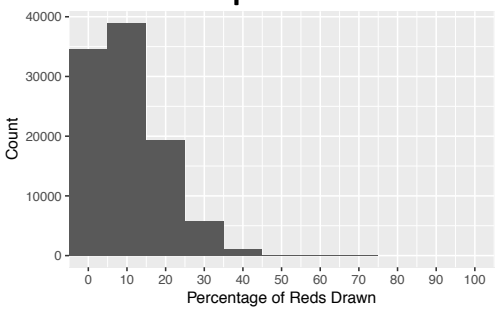
p = 0.3



p = 0.2



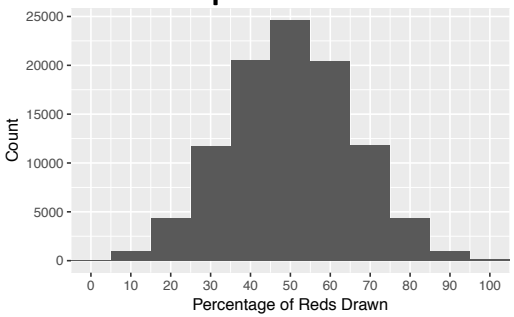
p = 0.1



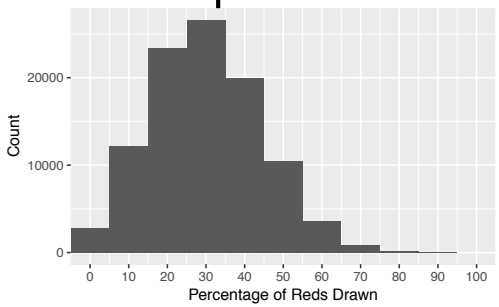
Remember some things we learned about proportions…

1. The farther the population proportion is from 0.5, the more skewed the sampling distribution is.

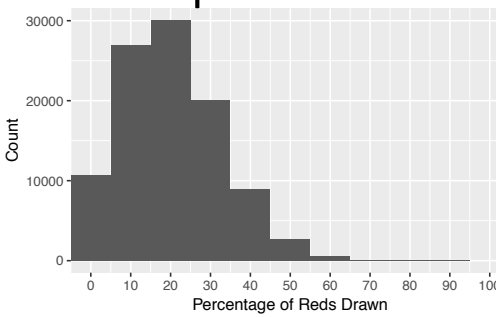
p = 0.5



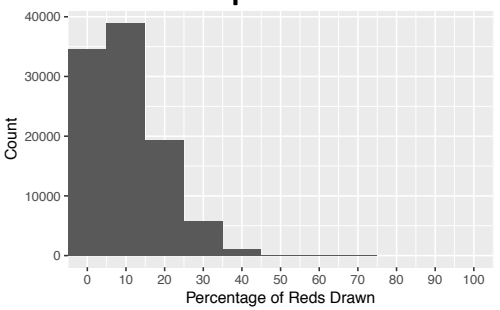
p = 0.3



p = 0.2

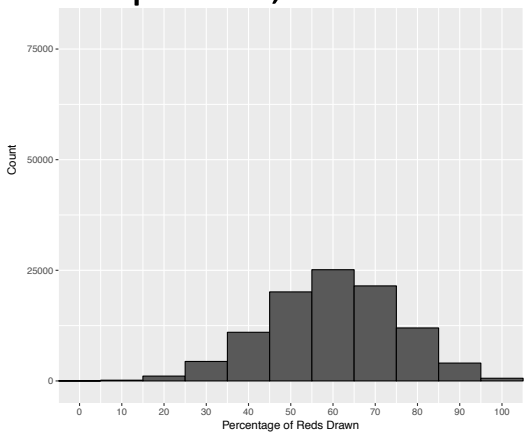


p = 0.1

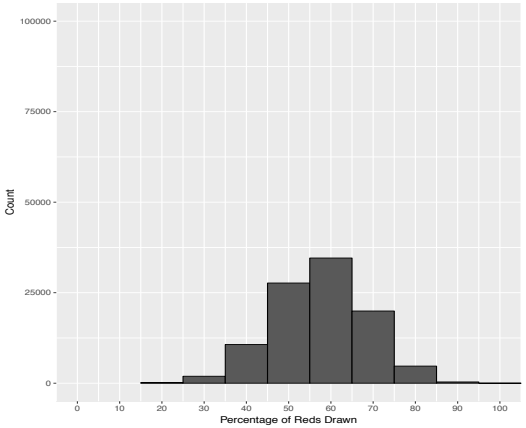


2. The larger your sample size, the narrower the sampling distribution is.

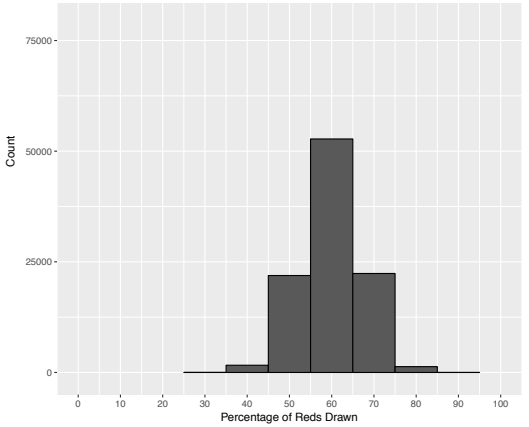
p = 0.6, n = 10



p = 0.6, n = 20

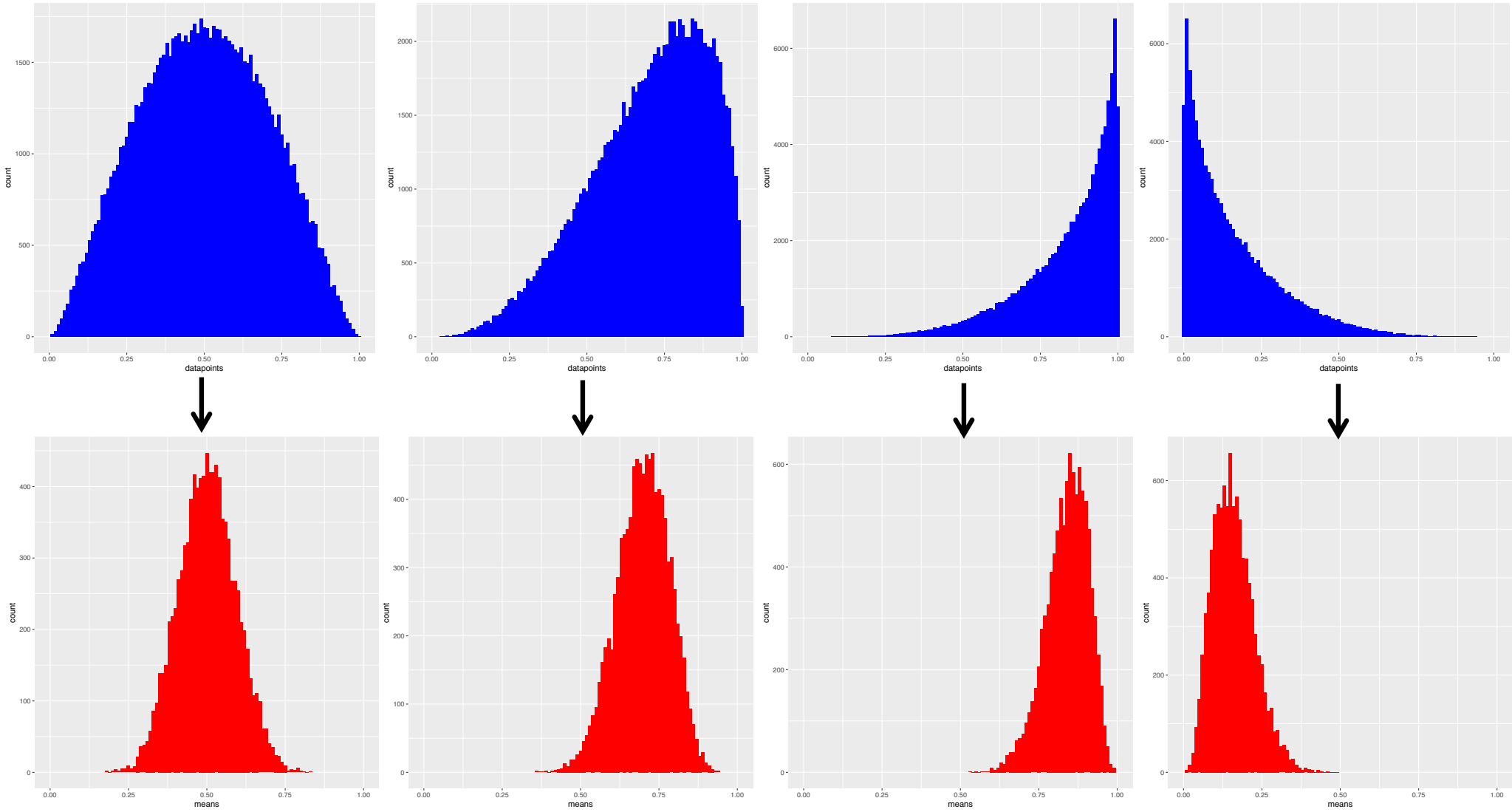


p = 0.6, n = 50



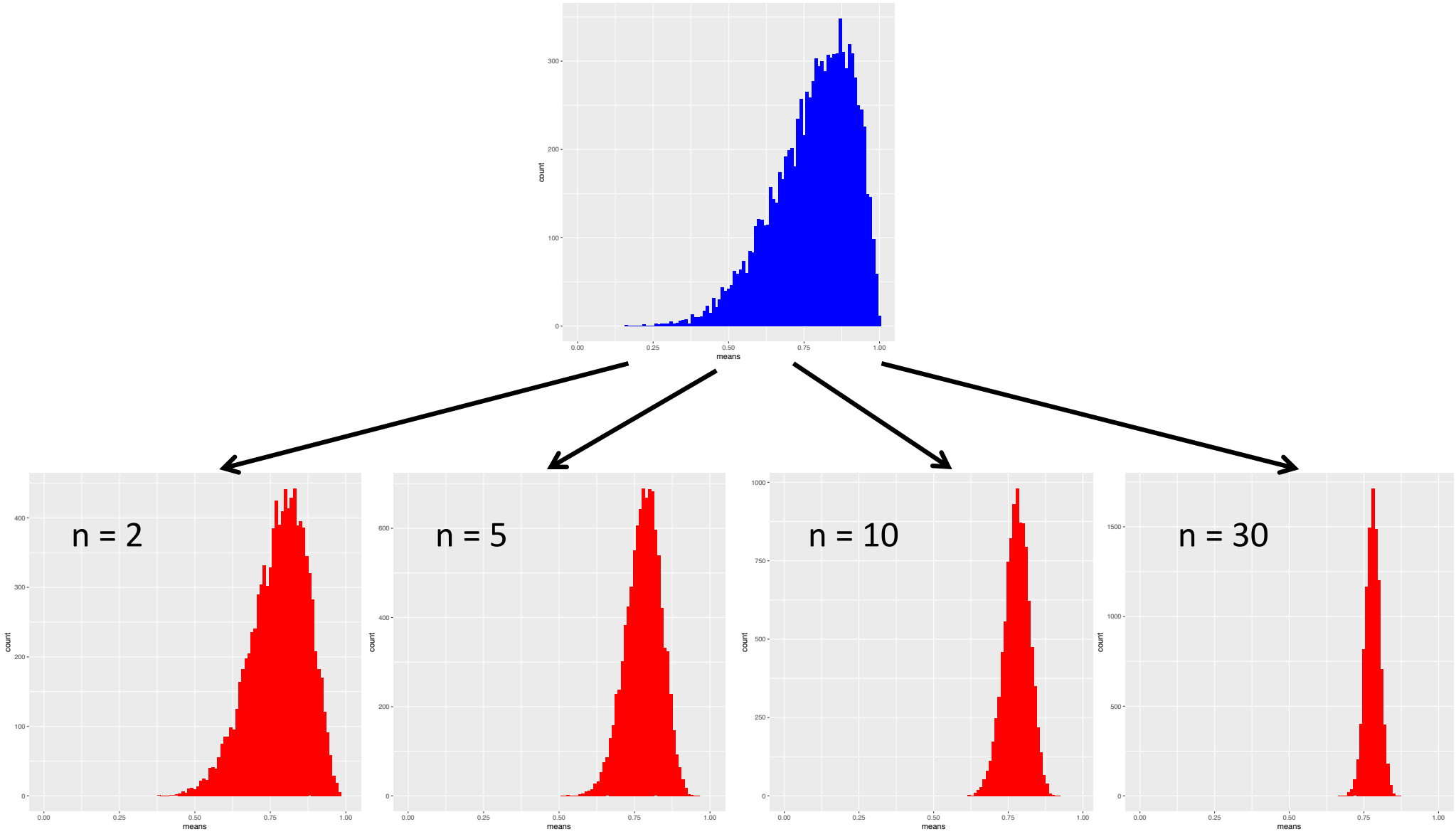
The same ideas apply to means…

1. The more skewed the underlying distribution, the more skewed the sampling distribution



The same ideas apply to means…

2. The larger your sample size, the narrower the sampling distribution is.



The Central Limit Theorem

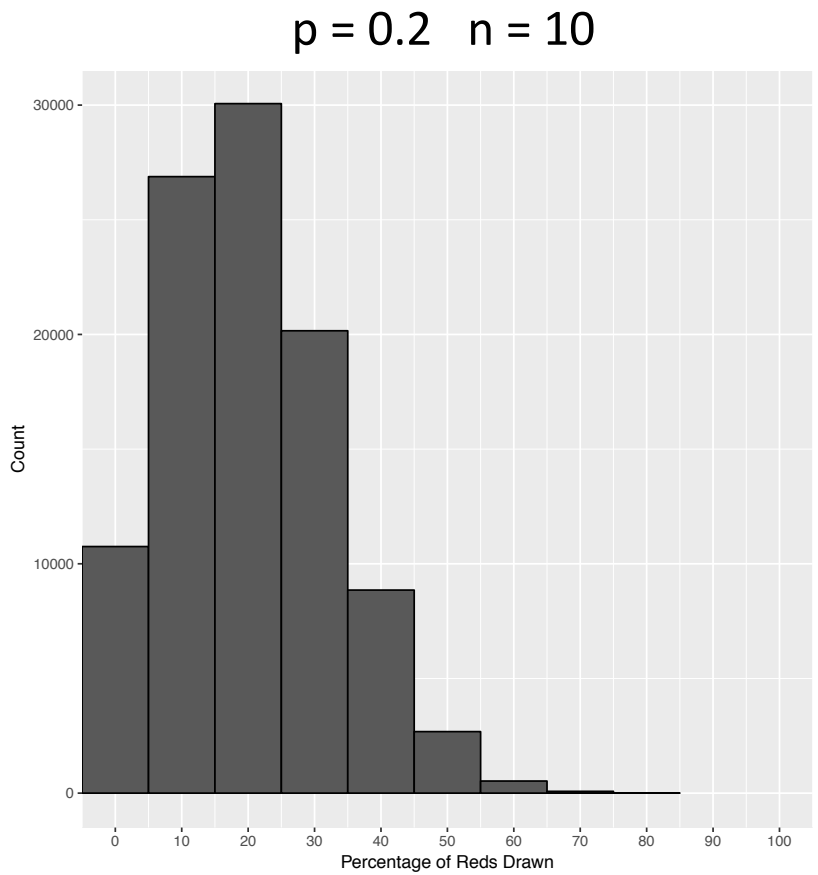
- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

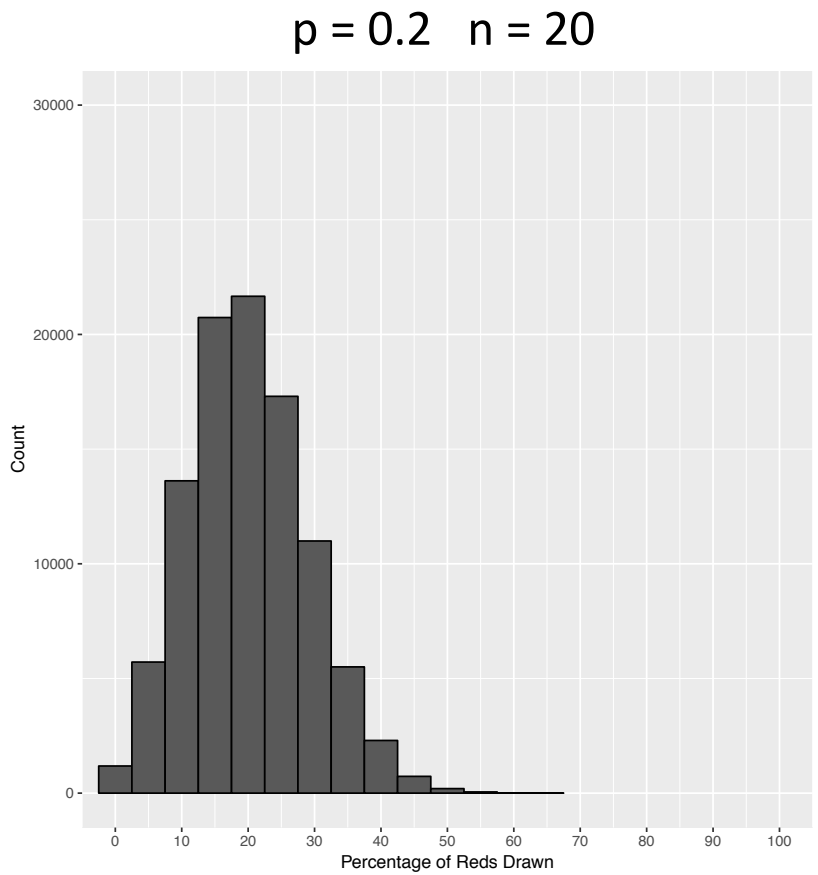
The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution



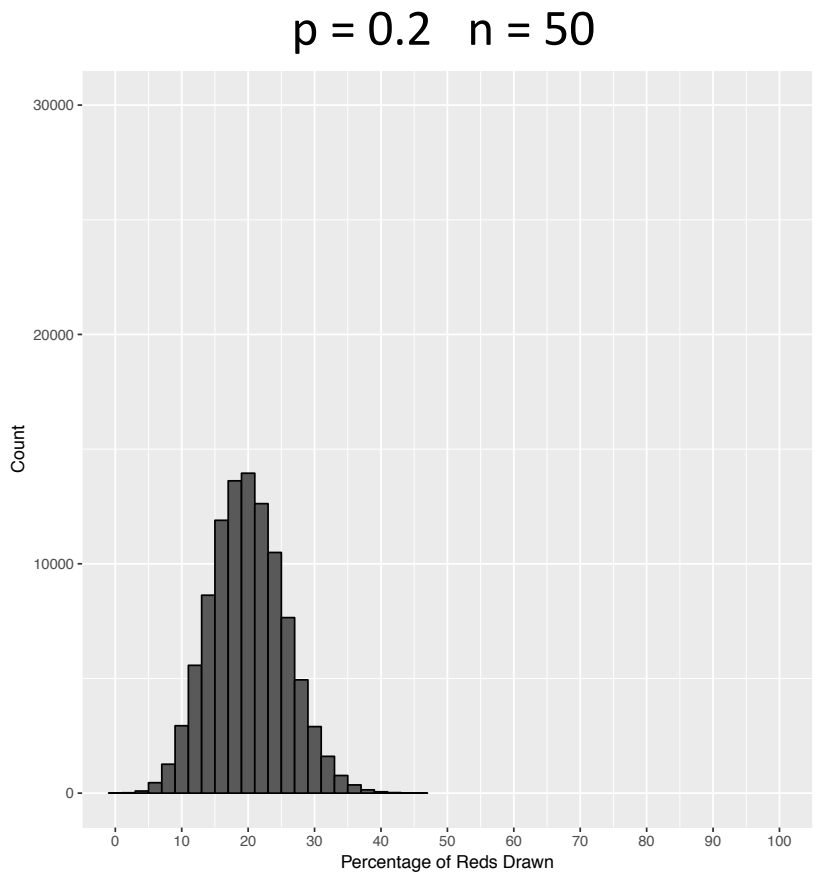
The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution



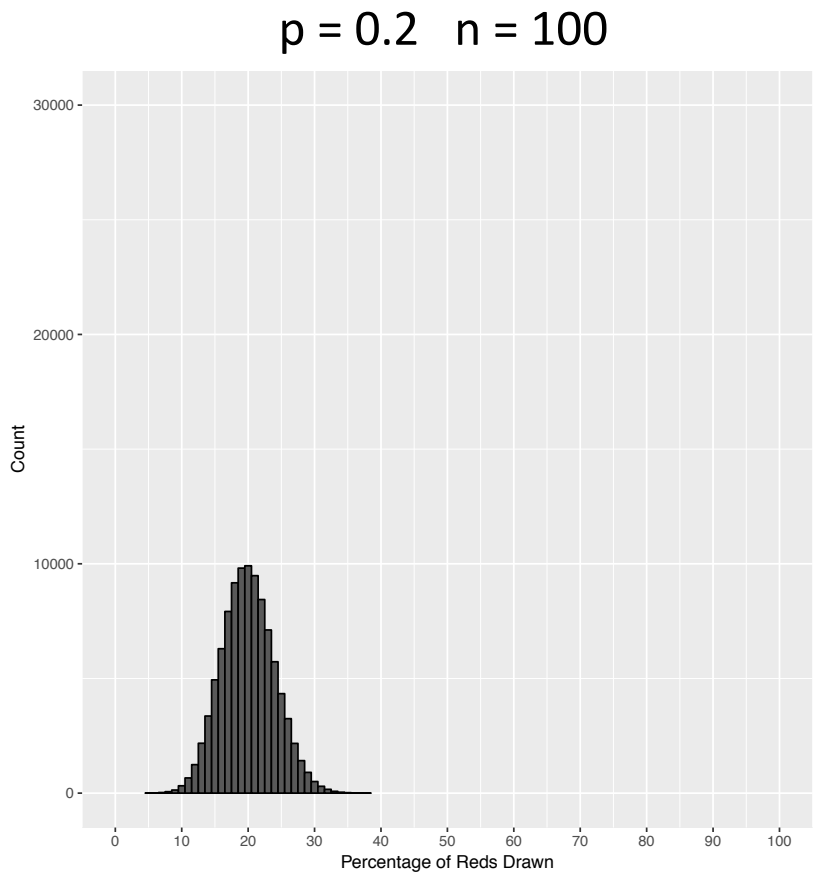
The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution



The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution



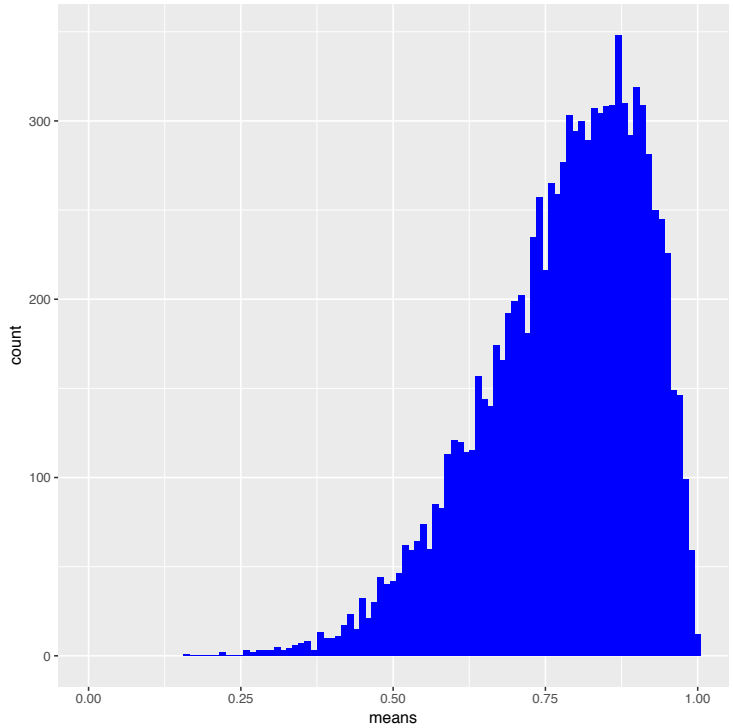
The Central Limit Theorem

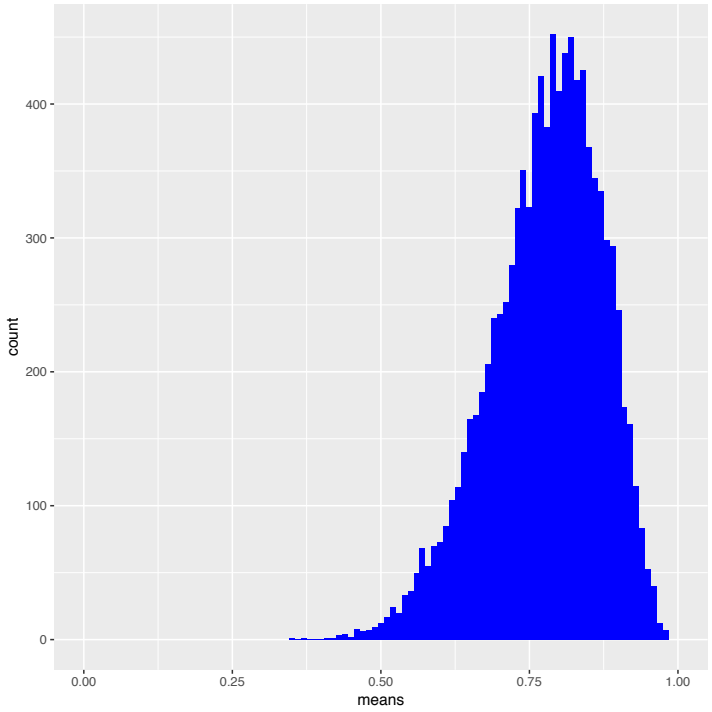
- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

n = 2

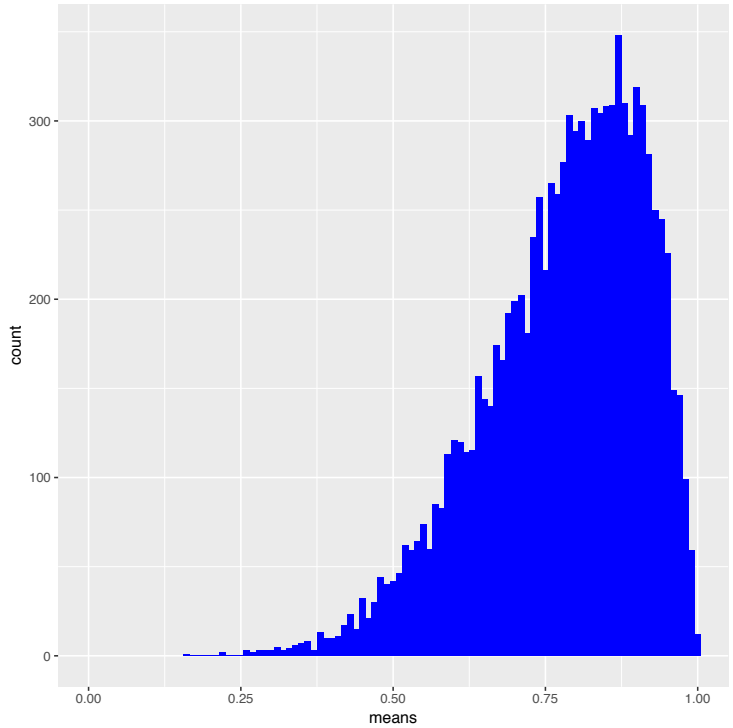


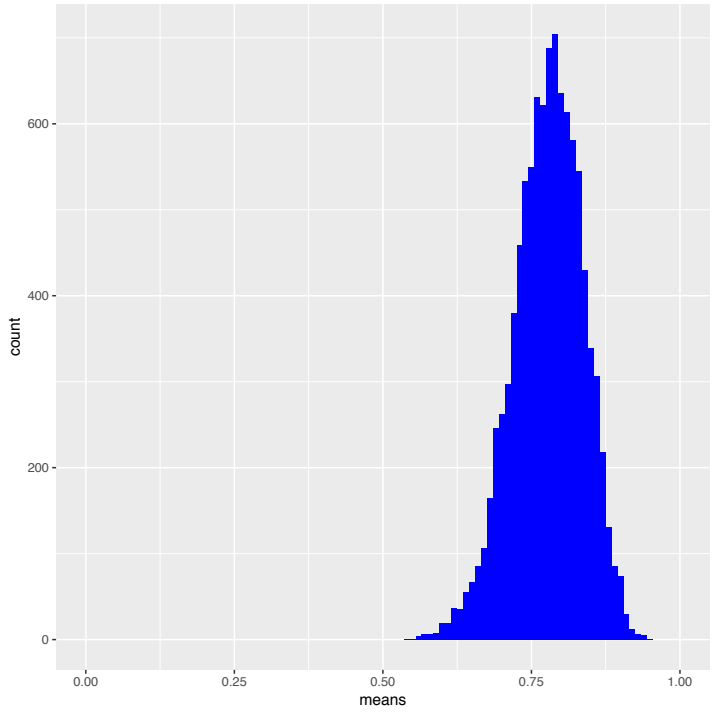


The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

n = 5

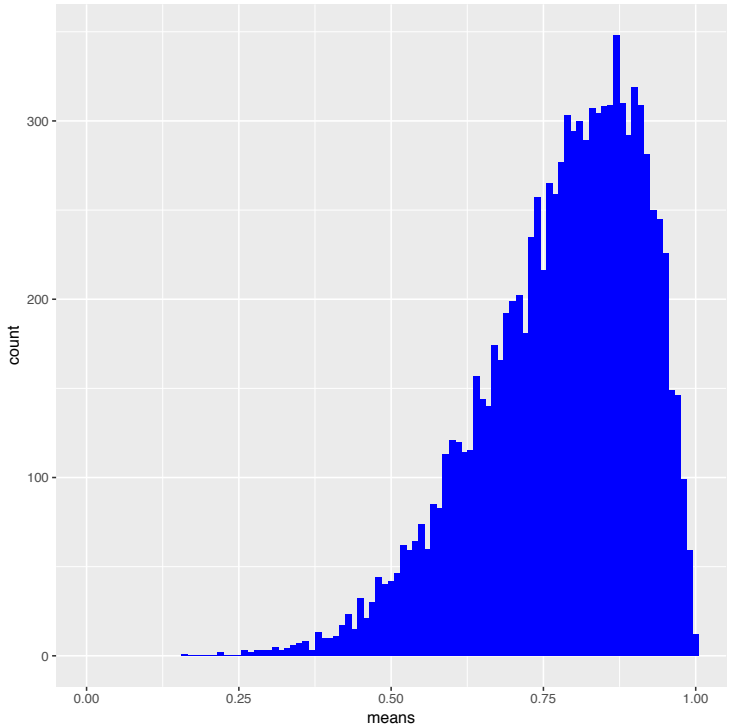


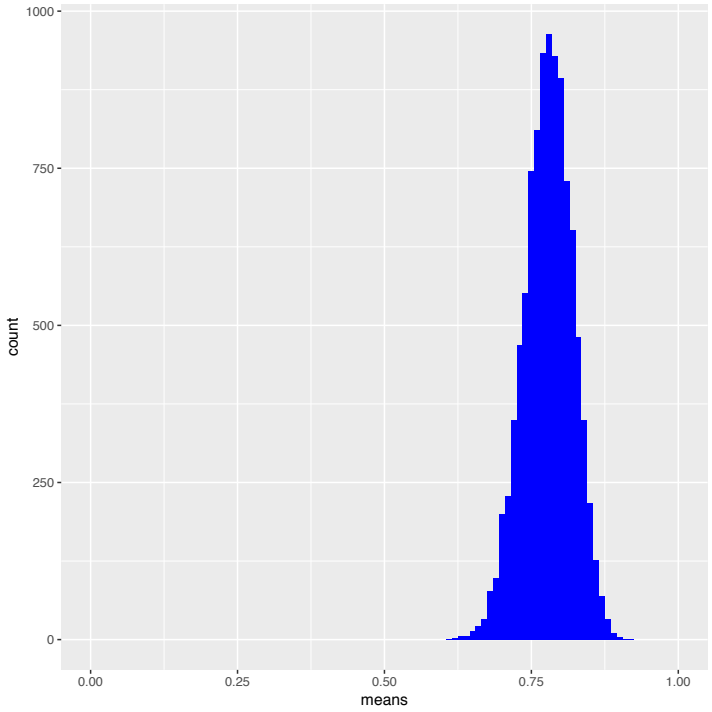


The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

n = 10

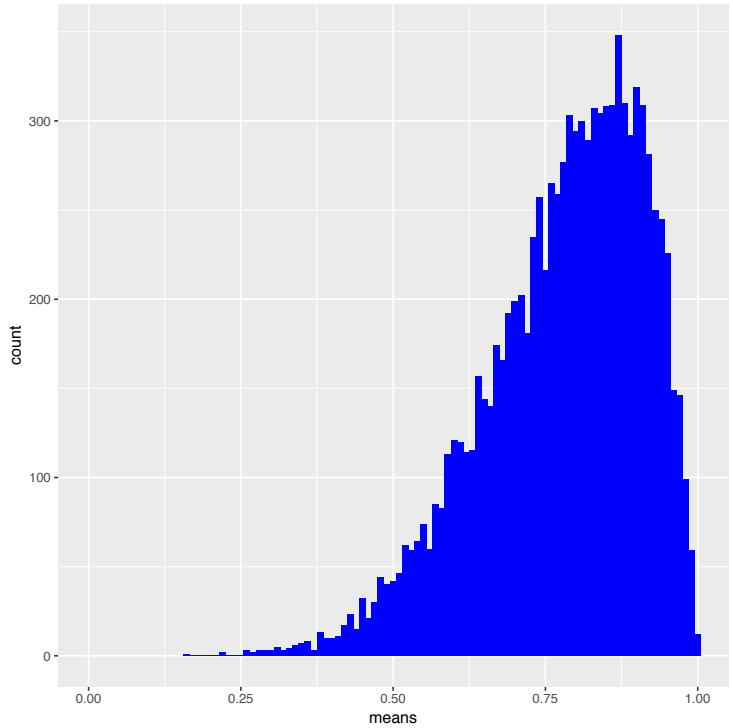


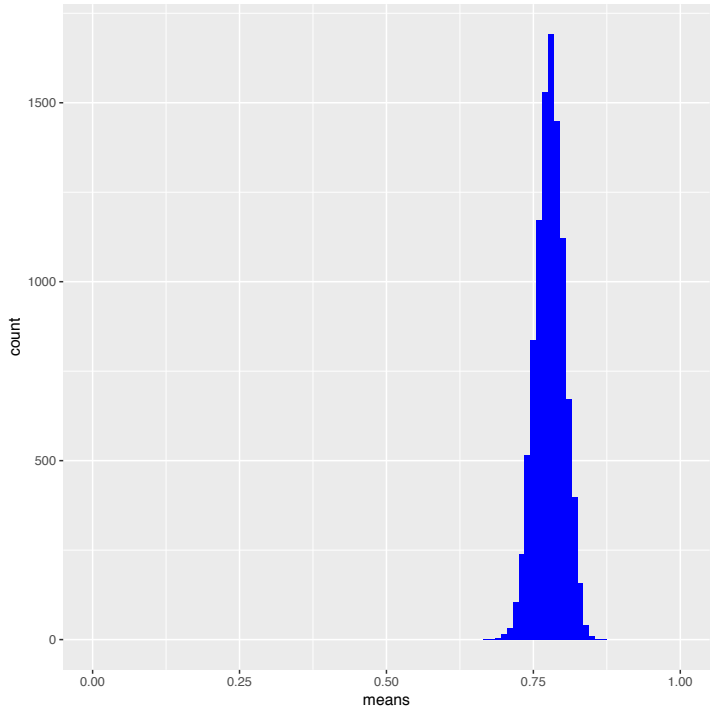


The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution

n = 30





The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution
- Why we care…
- For proportions we actually don't really care.
  - The exact shape of the sampling distribution depends on only two things:
  - The proportion in the underlying population (p) [We hypothesize this in our null hypothesis]
  - The sample size (n). [We know this]
  - With this info, the exact shape can be computed relatively easily (with modern computers).
    - In the olden days (30 years ago), this was more difficult to compute. So people pretended the sampling distribution was normal with mean p and SD (1-p)*p and based their inference off of that. We still do this frequently, though we no longer have to.

The Central Limit Theorem

- The CLT says: even if your population proportion is far from 0.5 (for proportions) or your underlying population distribution is highly skewed (for means), if you take a BIG enough sample, the sampling distribution will be a normal distribution
- Why we care…
- For means we still care.
- The exact shape of the sampling distribution for means depends on 4 things:
  - The mean in the underlying population [We hypothesize this in our null]
  - The SD in the underlying population [We make a "best" guess about this using prior info or info from our sample]
  - The sample size [We know this]
  - The exact shape of the distribution in the the underlying population
    - We don't know this. And it'd be hard to guess. Fortunately, the CLT says, it doesn't matter (as long as n is big). We can just assume the underlying distribution is normal, because even if it's not, it won't change the sampling distribution we ultimately get out (as long as n is big).