Nonparametric Statistics

1. An inference or model that does not use the probability distribution generating the data

2. Aspects of the probability dist'n may be known, but the complexity of the dist'n is unknown and is adaptive to the data (gets better with more data)

- Descriptive statistics and EDA are mostly nonparametric

- Semiparametric statistical inference: part of the model is parametric, part is nonparametric

<u>Ex</u>: $\underline{X_i \mid \mu_i} \sim \text{Normal}(\mu_i, 1)$.

$\underline{\mu_i \sim F}$ (arbitrary dist'n).

1. Empirical dist'n functions]
2. Bootstrap
3. Permutation methods — next class meeting
4. Goodness of fit
   - method of moments

<u>EDFs</u> (empirical distribution functions)

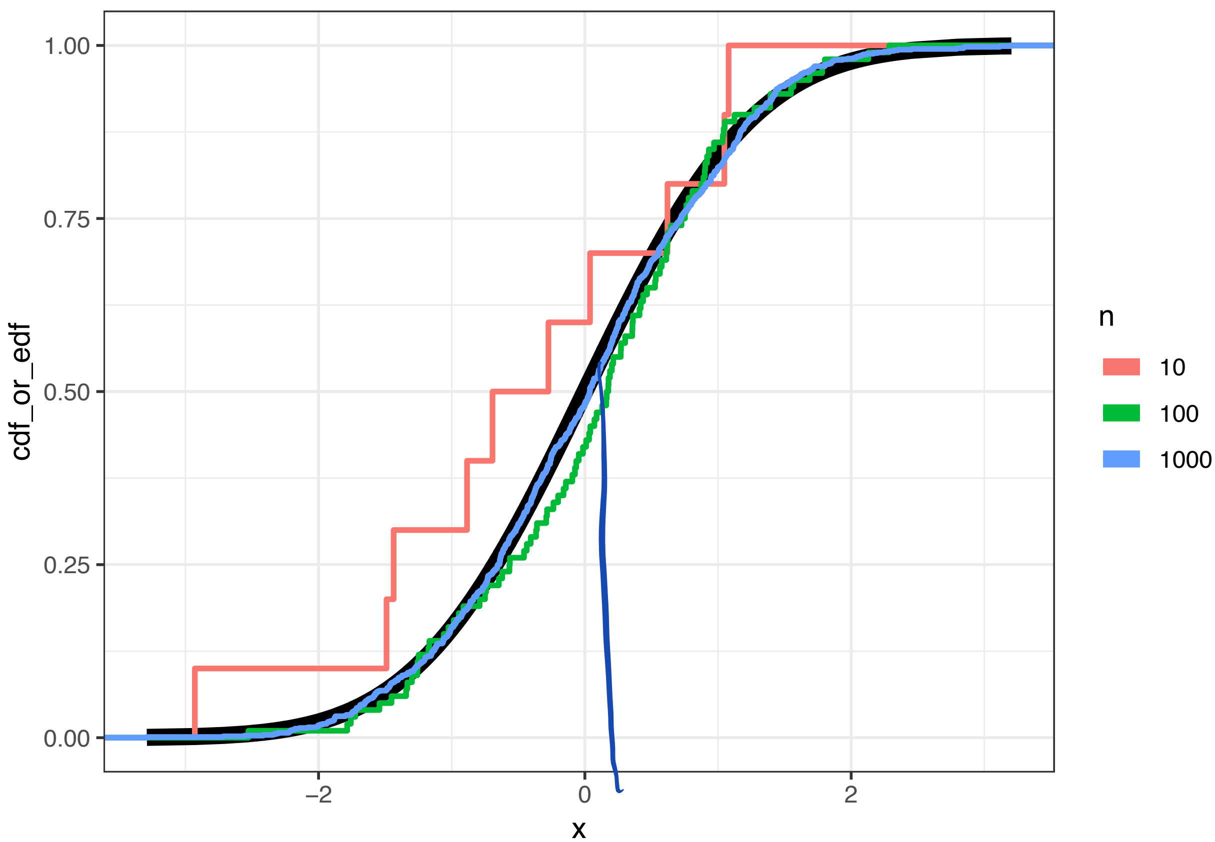
Say $X_1, X_2, \ldots, X_n \overset{iid}{\sim} \underline{F}$ (some distribution)

Let $\mathbb{1}(X_i \le y) = \begin{cases} 0 & \text{if } X_i > y \\ 1 & \text{if } X_i \le y \end{cases}$

Random variable: $\hat{F}_{\underline{X}}(y) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(X_i \le y)$

Observed variable: $F_{\underline{x}}(y) = \underline{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(x_i \le y)}$

EDF from Normal Data



Pointwise Convergence

$\hat{F}_{\underline{X}}(y) \longrightarrow F(y)$ w/ probability 1

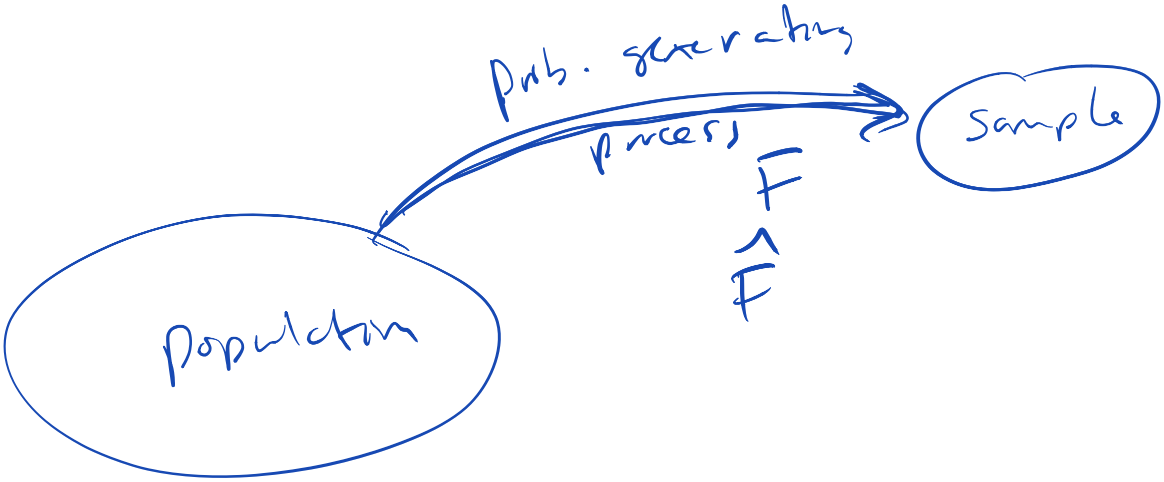
for each y

$n \to \infty$

Glivenko - Cantelli Theorem

$\left\{ \sup_{y \in \mathbb{R}} \left| \hat{F}_{\underline{X}}(y) - F(y) \right| \longrightarrow 0 \right.$

with probability 1



Statistical Functionals

A statistical functional $T(F)$ is any function of cdf, F. Examples:

- $\mu(F) = \int x \, dF(x)$ (general)

  $= \sum x f(x)$ (discrete)

  $= \int x f(x) \, dx$ (continuous)

- $\sigma^2(F) = \int (x - \mu(F))^2 \, dF(x)$

- $m(F) = F^{-1}(1/2)$

Plug-in Estimators of Statistical functionals from EDFs:

- $\hat{\mu} = \mu(\hat{F}) = \int x \, d\hat{F}(x)$
$$= \sum_{i=1}^{n} x_i \hat{f}(x_i)$$
$$= \sum_{i=1}^{n} x_i \frac{1}{n} \quad \text{(sample mean)}$$

- $\hat{\sigma}^2 = \sigma^2(\hat{F}) = \sum_{i=1}^{n} (x_i - \hat{\mu})^2 \frac{1}{n}$

- $\hat{m} = m(\hat{F}) = \hat{F}^{-1}(1/2)$

EDF CLT for a statistical functional

$$\left[\frac{T(F) - T(\hat{F})}{\hat{se}(T(\hat{F}))} \xrightarrow{n\to\infty} \text{Normal}(0,1) \right.$$

?

Linear Statistical Functional:

$$T(F) = \int a(x)\, dF(x)$$

$$\mathrm{Var}(T(\hat{F})) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{Var}(a(X_i))$$

$$= \frac{\mathrm{Var}(a(X))}{n}$$

$$se(T(\hat{F})) = \sqrt{\frac{\mathrm{Var}(a(X))}{n}}$$

$$\hat{se}(T(\hat{F})) = \sqrt{\frac{\mathrm{Var}_{\hat{F}}(a(X))}{n}}$$

Bootstrap

Basic idea: Use $\hat{F}$ (EDF) in place of F to get sampling distributions

Bootstrap Sample

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$

form $\hat{F}$ EDF

If I want n iid observations from $\hat{F}$, sample n observations with replacement from $\{X_1, X_2, \ldots, X_n\}$

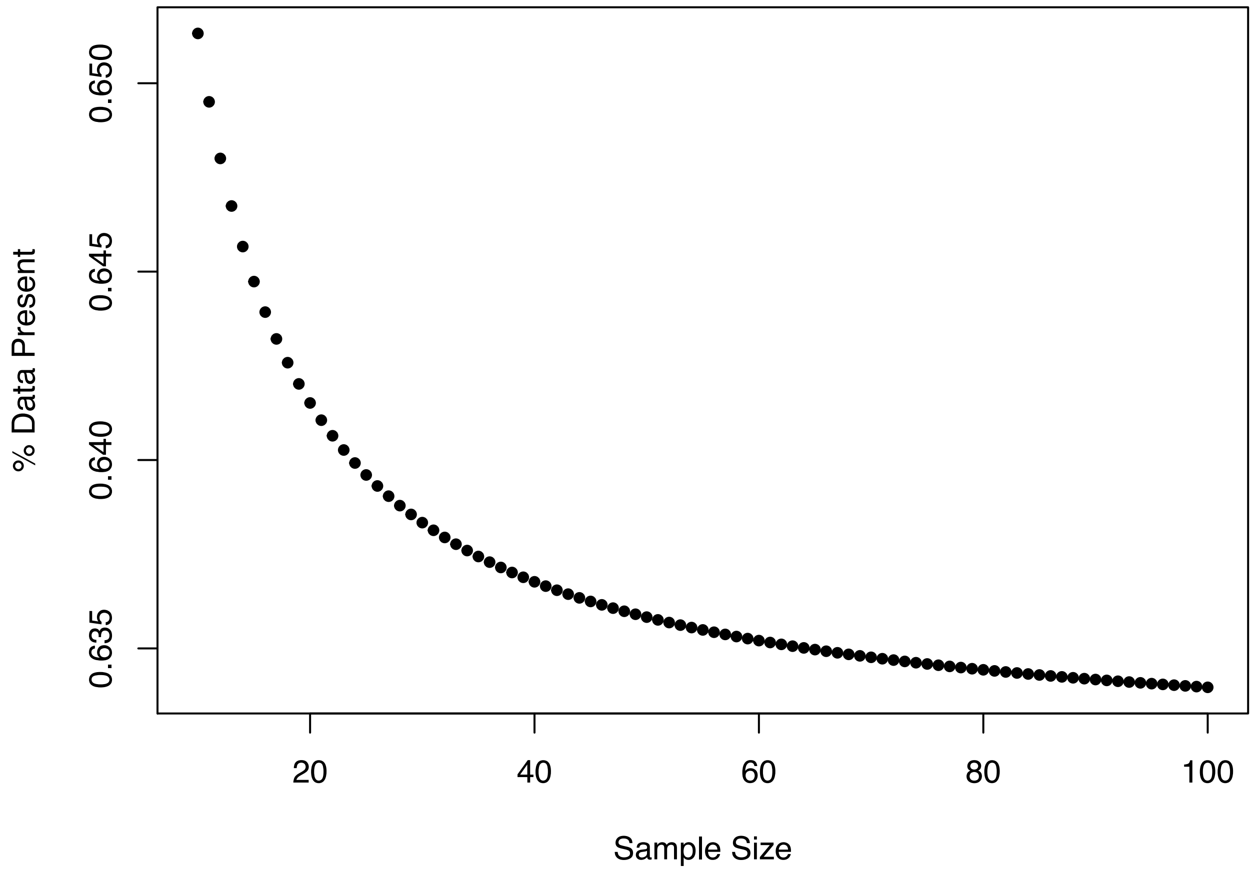
$\hat{F}$ puts probability $1/n$ on each X_i

Probability of not being sampled:

$$\left(1 - \frac{1}{n}\right)^n$$

P~~re~~ercentage of Data Present in a Bootstrap Sample

For a sample of size n, what percentage of the data is present in any given bootstrap sample?



Suppose we're interested in $\theta = T(F)$.

We estimate it by $\hat{\theta} = T(\hat{F}_{\underline{x}})$

<u>Idea</u>:

For $b = 1, 2, \dots, B$ we draw bootstrap data sets $x_1^{*(b)}, x_2^{*(b)}, \dots, x_n^{*(b)}$.

Example $n = 100$, $B = 10{,}000$

We can calculate the estimator $\hat{\theta}$ on each bootstrap sample:

$$\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots, \hat{\theta}^{*(B)}$$

Three ways to use bootstrap samples to get confidence intervals:

1. Percentile intervals
2. Pivotal intervals — 1st moment pivotal
3. studentized pivotal intervals — 2nd moment pivotal

$1-\alpha$ Bootstrap two-sided Confidence Intervals

1. Percentile interval:

Let $p^*_{\alpha/2}$ and $p^*_{1-\alpha/2}$ be

the $\alpha/2$ and $1-\alpha/2$ percentiles

of $\hat{\theta}^{*(1)}, \hat{\theta}^{*(2)}, \ldots, \hat{\theta}^{(B)}$.

CI is then:

$$(p^*_{\alpha/2}, p^*_{1-\alpha/2})$$

Aside Suppose $X_1, X_2, \ldots, X_n \sim Normal(0,1)$

EDF $\hat{F} \,\dot{\sim}\, Normal(\bar{X}, S^2)$

residuals

BS $X_i - \bar{X} \Longrightarrow \hat{F} \sim Normal(0, S^2)$

BS $\dfrac{X_i - \bar{X}}{\sqrt{S^2}} \Longrightarrow \hat{F} \sim Normal(0, 1)$

$z_{\alpha/2}$ $z_{1-\alpha/2}$

studentized residuals

(2) Pivotal interval (1st moment pivotal interval):

We calculate percentiles on $\hat{\theta}^{*(b)} - \hat{\theta}$. Call them q^*_α: They are bootstrap estimates of q_α, which are the α percentiles of $\hat{\theta} - \theta$

If we know q_α then the following is a $1-\alpha$ CI:

$$\left(\hat{\theta} - q_{1-\alpha/2}, \ \hat{\theta} - q_{\alpha/2} \right)$$

$$
\begin{aligned}
1-\alpha &= \Pr\left(q_{\alpha/2} \le \hat{\theta} - \theta \le q_{1-\alpha/2} \right) \\
&= \Pr\left(-q_{1-\alpha/2} \le \theta - \hat{\theta} \le -q_{\alpha/2} \right) \\
&= \Pr\left(\hat{\theta} - q_{1-\alpha/2} \le \theta \le \hat{\theta} - q_{\alpha/2} \right)
\end{aligned}
$$

Recall $\hat{\theta}^* - \hat{\theta}$ is approx. to $\hat{\theta} - \theta$

Suppose p^*_α is the α percentile of $\hat{\theta}^*$. Then $p^*_\alpha - \hat{\theta}$ is the approx. α percentile of $\hat{\theta}^* - \theta$

Therefore $p^*_\alpha - \hat{\theta}$ is the b.s. estimate of q_α. Plugging this into the above, we get the $(1-\alpha)$ CI is:

$$\left(2\hat{\theta} - p^*_{1-\alpha/2},\ 2\hat{\theta} - p^*_{\alpha/2}\right)$$

$$= \left(\hat{\theta} - q^*_{1-\alpha/2},\ \hat{\theta} - q^*_{\alpha/2}\right)$$

③ Studentized pivotal intervals (2nd moment pivotal)

The goal is to approximate the sampling dist'n of

$$\frac{\hat{\theta} - \theta}{se(\hat{\theta})}$$

Approximated by:

$$\frac{\hat{\theta}^* - \hat{\theta}}{\hat{se}(\hat{\theta}^*)}$$

Let z^*_α be the α percentile of

$$\left\{ \frac{\hat{\theta}^{*(1)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{*(1)})}, \frac{\hat{\theta}^{*(2)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{*(2)})}, \dots, \frac{\hat{\theta}^{*(B)} - \hat{\theta}}{\hat{se}(\hat{\theta}^{*(B)})} \right\}$$

Example: $\hat{\theta} = \bar{X}$

$\hat{\theta}^* = \bar{X}^*$

$\hat{se}(\hat{\theta}^*) = \dfrac{s^*}{\sqrt{n}}$

$$\frac{\hat{\theta}^* - \hat{\theta}}{\hat{se}(\hat{\theta}^*)} = \frac{\bar{X}^* - \bar{X}}{s^*/\sqrt{n}}$$

The $(1-\alpha)$ two-sided b.S. C.I. is

$$\left(\hat{\theta} - z^*_{1-\alpha/2}\,\hat{se}(\hat{\theta}),\ \hat{\theta} - z^*_{\alpha/2}\,\hat{se}(\hat{\theta})\right)$$

replacing Normal(0,1) percentiles

How do we get $\hat{se}(\hat{\theta})$ in nonparametric setting?

$$\hat{se}(\hat{\theta}) = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\hat{\theta}^{*(b)} - \frac{1}{B}\sum_{k=1}^{B}\hat{\theta}^{*(k)}\right)^2}$$

But how to get $\hat{se}(\hat{\theta}^{*(b)})$???

$Exp(\lambda)$ $E[X] = 1/\lambda$ $= 2$

$\lambda = 1/2$

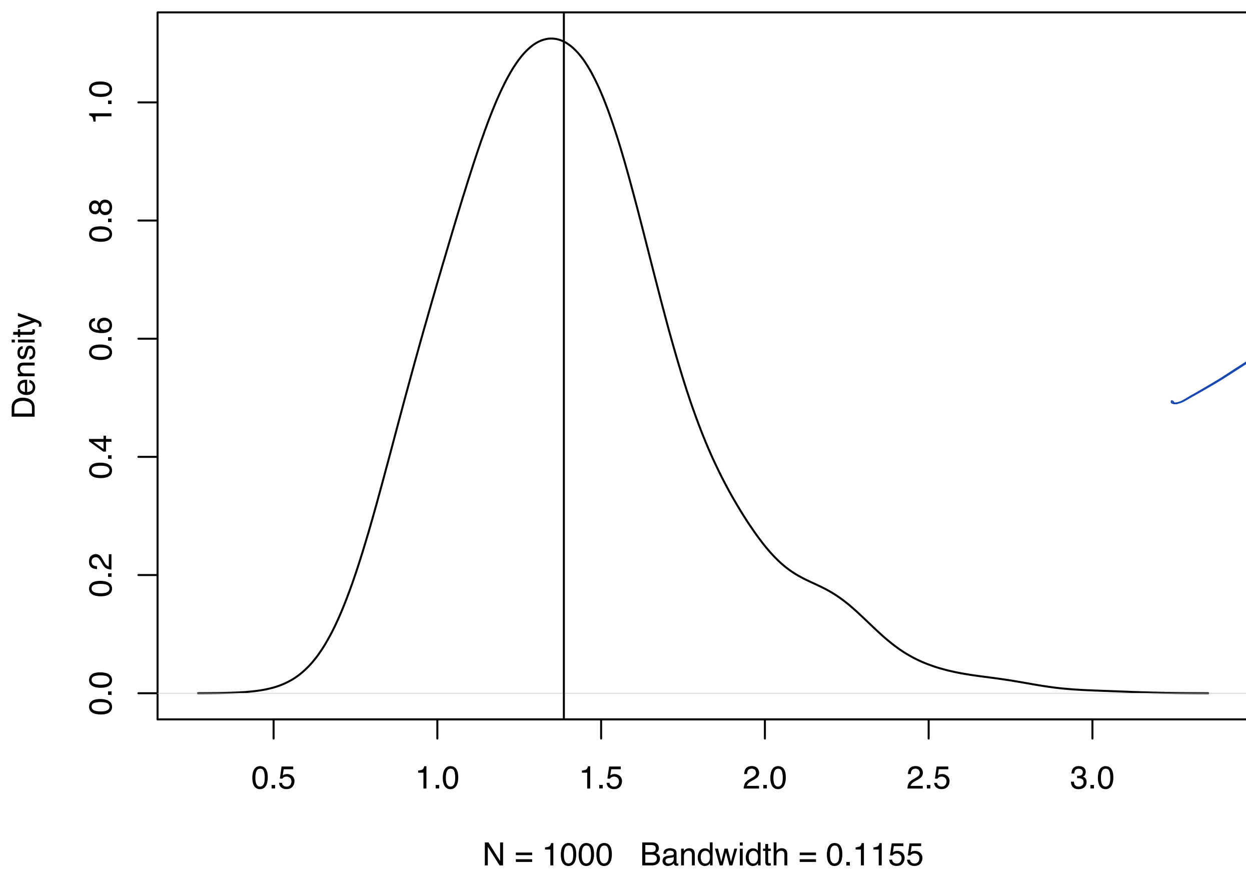
Example: Bootstrap on Exponential Data

In the homework, you will be performing a bootstrap t-test of the mean and a bootstrap percentile CI of the median for the following Exponential(λ) data:

```
> set.seed(1111)
> pop_mean <- 2
> X <- matrix(rexp(1000*30, rate=1/pop_mean), nrow=1000, ncol=30)
```

Let's construct a pivotal bootstrap CI of the median here instead.

```
> # population median 2*log(2)
> pop_med <- qexp(0.5, rate=1/pop_mean); pop_med
[1] 1.386294
>
> obs_meds <- apply(X, 1, median)
> plot(density(obs_meds, adj=1.5), main=" "); abline(v=pop_med)
```



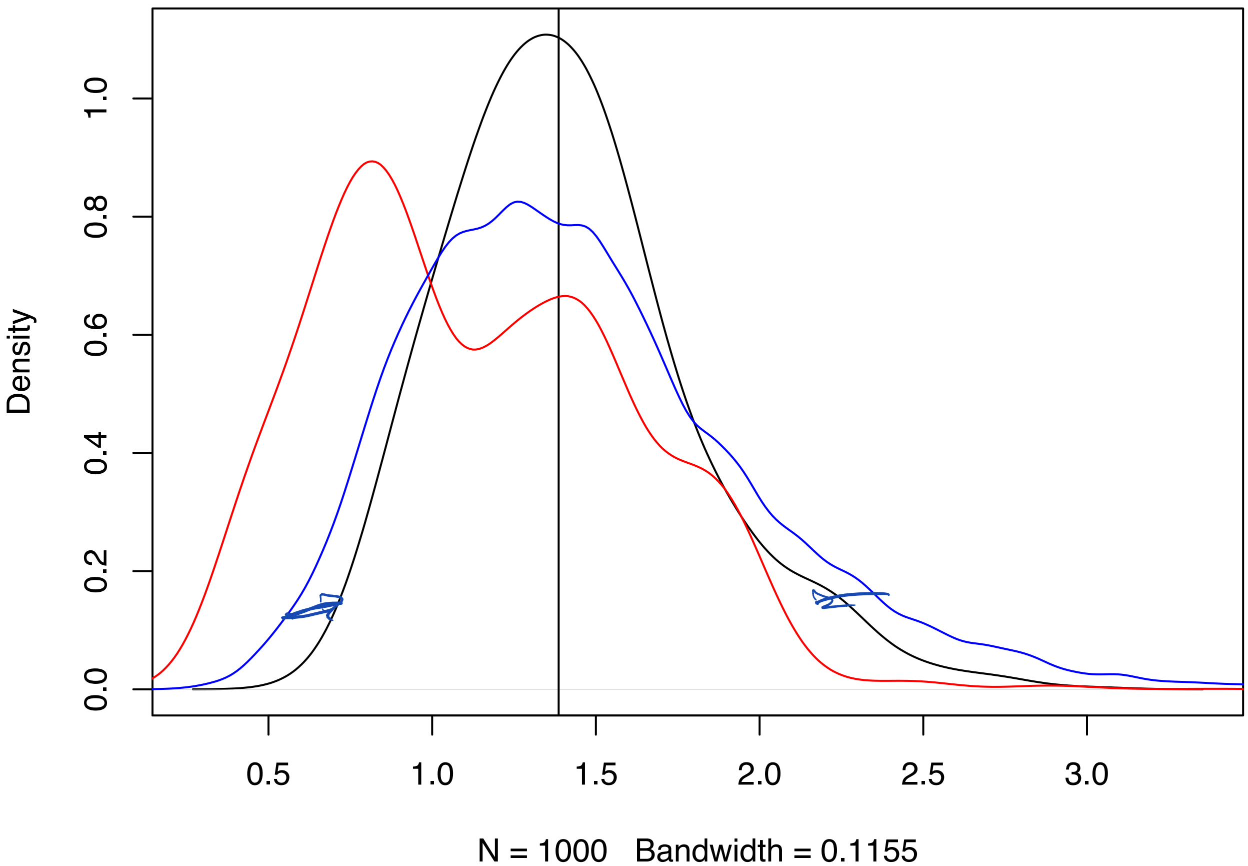
Some embarrassingly inefficient code to calculate bootstrap medians.

```
> B <- 1000
> boot_meds <- matrix(0, nrow=1000, ncol=B)
>
> for(b in 1:B) {
```

```
+   idx <- sample(1:30, replace=TRUE)
+   boot_meds[,b] <- apply(X[,idx], 1, median)
+ }
```

Plot the bootstrap medians.

```
> plot(density(obs_meds, adj=1.5), main=" "); abline(v=pop_med)
> lines(density(as.vector(boot_meds[1:4,]), adj=1.5), col="red")
> lines(density(as.vector(boot_meds), adj=1.5), col="blue")
```

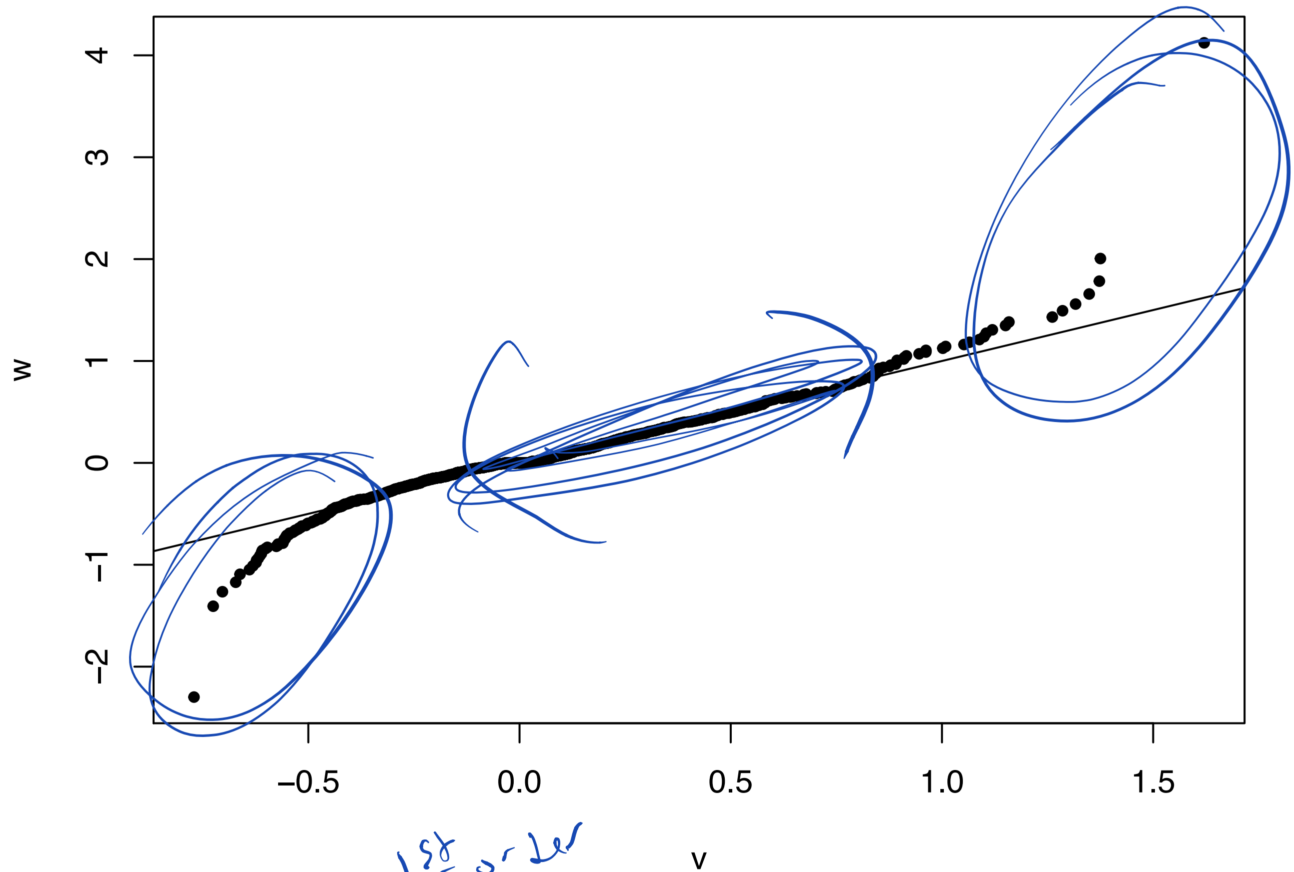


Compare sampling distribution of $\hat{\theta} - \theta$ to $\hat{\theta}^* - \hat{\theta}$.

```
> v <- obs_meds - pop_med
> w <- as.vector(boot_meds - obs_meds)
> qqplot(v, w, pch=20); abline(0,1)
```

$\hat{\theta} - \theta$

$\sim \hat{\theta}^* - \hat{\theta}$



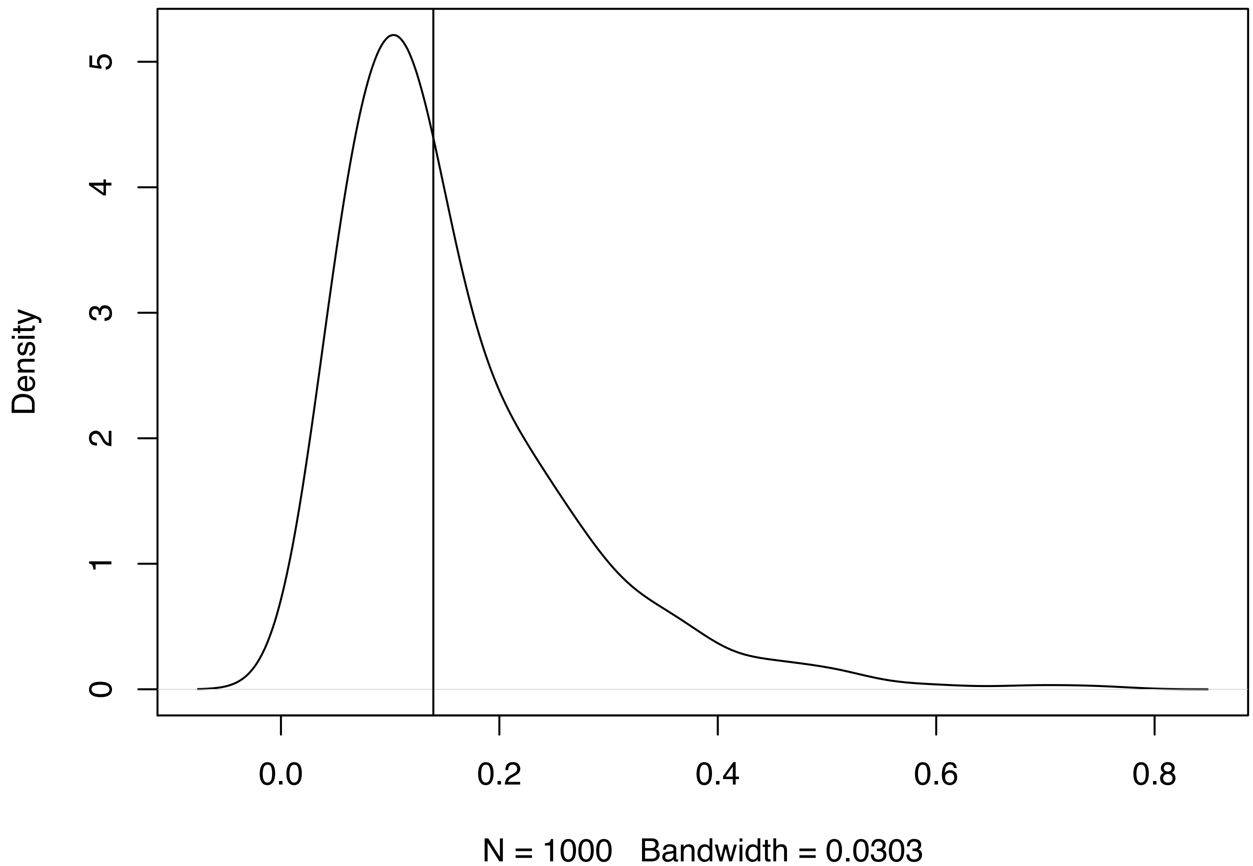
1st order

Does a 95% bootstrap pivotal interval provide coverage?

```
> ci_lower <- apply(boot_meds, 1, quantile, probs=0.975)
> ci_upper <- apply(boot_meds, 1, quantile, probs=0.025)
>
> ci_lower <- 2*obs_meds - ci_lower
> ci_upper <- 2*obs_meds - ci_upper
>
> ci_lower[1]; ci_upper[1]
[1] 0.8958224
[1] 2.113859
>
> cover <- (pop_med >= ci_lower) & (pop_med <= ci_upper)
> mean(cover)
[1] 0.809
>
> # :-(
```

Let's check the bootstrap variances.

```
> sampling_var <- var(obs_meds)
> boot_var <- apply(boot_meds, 1, var)
> plot(density(boot_var, adj=1.5), main=" ")
> abline(v=sampling_var)
```



We repeated this simulation over a range of n and B.

n	B	*coverage*	*avg CI width*
1e+02	1000	0.868	0.7805404
1e+02	2000	0.872	0.7882278
1e+02	4000	0.865	0.7852837
1e+02	8000	0.883	0.7817222
1e+03	1000	0.923	0.2465840
1e+03	2000	0.909	0.2477463
1e+03	4000	0.915	0.2475550
1e+03	8000	0.923	0.2458167
1e+04	1000	0.935	0.0781421
1e+04	2000	0.937	0.0784541
1e+04	4000	0.942	0.0784559
1e+04	8000	0.948	0.0785591
1e+05	1000	0.949	0.0246918
1e+05	2000	0.942	0.0246938

Goodness of Fit Methods

We don't know the dist'n of the data, but we'd like to test or assess its fit to a known distribution

1. Chi-square GoF
2. KS Test
3. Method of moments

Chi-square GoF

$X_1, X_2, \ldots, X_n \overset{iid}{\sim} F$

Test: $H_0: F \in \{F_\theta : \theta \in \Theta\}$

H_1: not H_0

Divide the support of $\{F_\theta : \theta \in \Theta\}$ in K bins $I_1, I_2, \ldots, I_K$

Example : Normal (μ, σ^2)

$$(-\infty, -10), (-10, -9] \quad \cdots \quad (9, 10), (10, \infty)$$
$$I_1 \qquad I_2 \qquad \cdots \qquad I_{K-1} \qquad I_K$$

For $j = 1, 2, \ldots, K$ calculate

$$\underline{q_j(\theta)} = \int_{I_j} \underline{dF_\theta(x)}$$

Suppose observe data $x_1, x_2, \ldots, x_n$.

Let n_j be the number of data points in interval I_j.

Let $\tilde{\theta}$ be the value of θ that is the MLE of :

$$\prod_{j=1}^{K} q_j(\theta)^{n_j}$$

Form GoF statistic:

$$S(\underline{x}) = \sum_{j=1}^{K} \frac{(n_j - n q_j(\tilde{\theta}))^2}{n q_j(\tilde{\theta})}$$

$n q_j(\tilde{\theta})$ is the expected number of observations in I_j with parameter values $\tilde{\theta}$

when H_0 is true, $S(\underline{x})$ has a χ^2_v where $v = K - \dim(\theta) - 1$

p-value $= \Pr(S(\underline{X}^*) \geq s(\underline{x}))$
where $S(X^*) \sim \chi^2_v$.

Goodness of Fit Example: Hardy-Weinberg

Suppose at your favorite SNP, we observe genotypes from 100 randomly sampled individuals as follows:

AA	AT	TT
28	60	12

If we code these genotypes as 0, 1, 2, testing for Hardy-Weinberg equilibrium is equivalent to testing whether $X_1, X_2, \ldots, X_{100} \overset{\text{iid}}{\sim}$ Binomial$(2, \theta)$ for some unknown allele frequency of T, θ.

The parameter dimension is such that $d = 1$. We will also set $k = 3$, where each bin is a genotype. Therefore, we have $n_1 = 28$, $n_2 = 60$, and $n_3 = 12$. Also,

$$q_1(\theta) = (1-\theta)^2, \quad q_2(\theta) = 2\theta(1-\theta), \quad q_3(\theta) = \theta^2.$$

Forming the multinomial likelihood under these bin probabilities, we find $\tilde{\theta} = (n_2 + 2n_3)/(2n)$. The degrees of freedom of the χ^2_v null distribution is $v = k - d - 1 = 3 - 1 - 1 = 1$.

Let's carry out the test in R.

```
> n <- 100
> nj <- c(28, 60, 12)
>
> # parameter estimates
> theta <- (nj[2] + 2*nj[3])/(2*n)
> qj <- c((1-theta)^2, 2*theta*(1-theta), theta^2)
>
> # gof statistic
> s <- sum((nj - n*qj)^2 / (n*qj))
>
> # p-value
> 1-pchisq(s, df=1)
[1] 0.02059811
```

Kolmogorov-Smirnov Test

(1) Form EDF $\hat{F}$

(2) Parametric F_θ (θ known)

(3) Form statistic:

$$D(\underline{x}) = \max_y \left| \hat{F}_{\underline{x}}(y) - F_\theta(y) \right|$$

Null distribution of $D(\underline{x})$ is known, based on Brownian bridge.

$H_0: F = F_\theta$ vs. $H_1: F \neq F_\theta$

Two-sample KS-test

$X_1, \ldots, X_n \sim F_X$

$Y_1, \ldots, Y_m \sim F_Y$

$H_0: F_X = F_Y$ vs $H_1: F_X \neq F_Y$

$$D(\underline{x}, \underline{y}) = \max_{z} \left| \hat{F}_{\underline{x}}(z) - \hat{F}_{\underline{y}}(z) \right|$$

When H_0 is true, one can calculate the dist'n of

$D(\underline{x}, \underline{y})$.

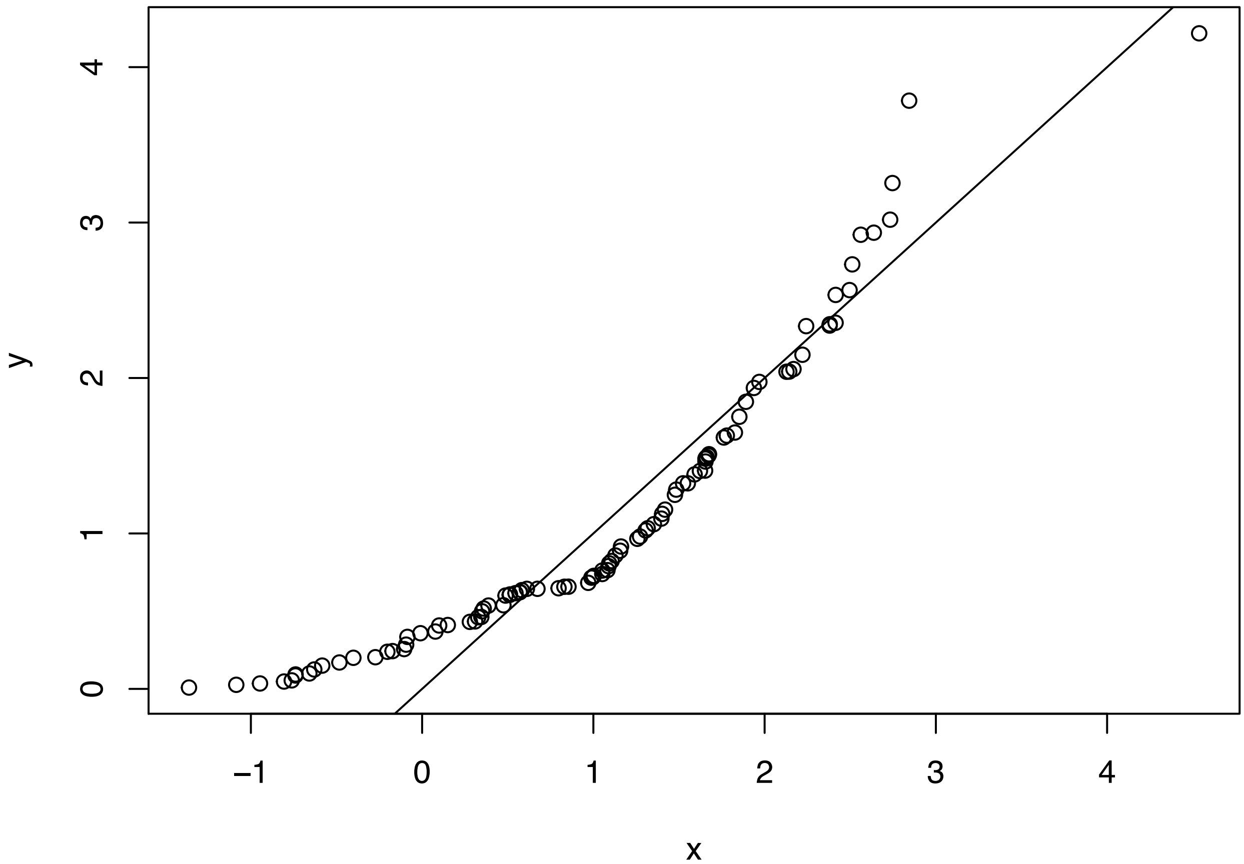
KS Test Example: Exponential vs Normal

```
ks.test(x, y, ...,
        alternative = c("two.sided", "less", "greater"),
        exact = NULL)
```

Two sample KS test.

```
> x <- rnorm(100, mean=1)
> y <- rexp(100, rate=1)
> wilcox.test(x, y)

    Wilcoxon rank sum test with continuity correction

data:  x and y
W = 5021, p-value = 0.9601
alternative hypothesis: true location shift is not equal to 0
> ks.test(x, y)

    Two-sample Kolmogorov-Smirnov test

data:  x and y
D = 0.19, p-value = 0.0541
alternative hypothesis: two-sided
```

```
> qqplot(x, y); abline(0,1)
```



One sample KS tests.

```
> ks.test(x=x, y="pnorm")

	One-sample Kolmogorov-Smirnov test

data:  x
D = 0.41398, p-value = 2.554e-15
alternative hypothesis: two-sided
>
> ks.test(x=x, y="pnorm", mean=1)

	One-sample Kolmogorov-Smirnov test

data:  x
D = 0.068035, p-value = 0.7436
alternative hypothesis: two-sided
```

Standardize (mean center, sd scale) the observations before comparing to a Normal(0,1) distribution.

```
> ks.test(x=((x-mean(x))/sd(x)), y="pnorm")
```

```
    One-sample Kolmogorov-Smirnov test

data:  ((x - mean(x))/sd(x))
D = 0.05896, p-value = 0.8778
alternative hypothesis: two-sided
>
> ks.test(x=((y-mean(y))/sd(y)), y="pnorm")

    One-sample Kolmogorov-Smirnov test

data:  ((y - mean(y))/sd(y))
D = 0.14439, p-value = 0.03092
alternative hypothesis: two-sided
```